iConference 2013 Organizers

Program Chairs

Honorary Conference Co-Chairs
Elizabeth D. Liddy, Syracuse University
Herman L. Totten, University of North Texas

Conference Chair
William E. Moen, University of North Texas

Program Co-Chairs
Kevin Crowston, Syracuse University
Martin Halbert, University of North Texas

Papers and Notes Co-Chairs
Kevin Crowston, Syracuse University
Martin Halbert, University of North Texas

Posters Chair
Catherine Blake, University of Illinois

Workshops Chair
John Carlo Bertot, University of Maryland, College Park
Paul T. Jaeger, University of Maryland, College Park

Alternative Events Co-Chairs
Marcia A. Mardis, Florida State University
Maria L. Souden, University College Dublin

Research Paper Development Roundtable Chair
Martin B.H. Weiss, University of Pittsburgh.

Doctoral Colloquium Co-Chairs
Hamid Ekbia, Indiana University
Karen Fisher, University of Washington
Jens-Erik Mai, Royal School of Library and Information Science

Early Career Colloquium Chair
Steven B. Sawyer, Syracuse University
Kristin R. Eschenfelder, University of Wisconsin-Madison

Proceedings Co-Chairs
Linda Schamber, University of North Texas
Oksana L. Zavalina, University of North Texas

ConfTool System Coordinator
Yunfei Du, University of North Texas

Conference Coordinator
Clark Heideger, iCaucus

Local Arrangements Committee Chair
Laura Venhaus, University of North Texas

Program Committee

Eileen Abels, Drexel University
Theresa Anderson, University of Technology, Sydney
Rosa Arriaga, Georgia Institute of Technology
Nicholas Belkin, Rutgers, The State University of New Jersey
John Carlo Bertot, University of Maryland
Randolph Bias, University of Texas
Wade Bishop, University of Kentucky
Catherine Blake, University of Illinois
Pia Borlund, Royal School of Library and Information Science Copenhagen
Peter Brusilovsky, University of Pittsburgh
Joy Davidson, Glasgow University
Robert Deng, Singapore Management University
Yunfei Du, University of North Texas
Hamid Ekbia, Indiana University
Pedro Ferreira, Carnegie Mellon
Karen Fisher, University of Washington
Andrew Flinn, University College London
Fred Fonseca, Pennsylvania State University
Maria Gade, Humboldt-Universität zu Berlin
Steve Howard, University of Melbourne
Paul T. Jaeger, University of Maryland
Heikki Keskustalo, University of Tampere

Anita Komlodi, University of Maryland, Baltimore County
Cory Knobel, University of California, Irvine
Elizabeth Liddy, Syracuse University
Jens-Erik Mai, Royal School of Library and Information Science
Marcia A. Mardis, The Florida State University
Eden Medina, Indiana University
Eric Meyers, University of British Columbia
William Moen, University of North Texas
Karine Nahon, University of Washington
Arcot Rajasekar (Raja), University of North Carolina at Chapel Hill
Lionel Robert, University of Michigan
Howard Rosenbaum, Indiana University
Steve Sawyer, Syracuse University
Linda Schamber, University of North Texas
Barbara Schultz-Jones, University of North Texas
Maria Souden, University College Dublin
Mega Subramanian, University of Maryland
Elaine Toms, University of Sheffield
Herman Totten, University of North Texas
Martin H.B. Weiss, University of Pittsburgh
Judith Wusteman, University College Dublin
Iris Xie, University of Wisconsin-Milwaukee
Oksana L. Zavalina, University of North Texas
Doctoral Dissertation Award Coordinating Committee

David Hendry, University of Washington
Lynne Howarth, University of Toronto
Steven B. Sawyer, Syracuse University

Social Media Expo Organizing Committee

Shelly D. Farnham, Microsoft Research
Eytan Adar, University of Michigan
Jamie Callan, Carnegie Mellon University
Anthony J. Rotolo, Syracuse University

Local Arrangements Committee*

Laura Venhaus (Chair)
Angelique Alexander
Jurhee Curtis
Brenda Fanara
Kathryn Masten

Reviewers

Eileen Abels, Drexel University
Rafa Absar, University of British Columbia
Jae-wook Ahn, Drexel University
Dharma Akmon, University of Michigan
Larry Alexander, Drexel University
Rosa Arriaga, Georgia Institute of Technology
Paavo Arvola, University of Tampere
Jung Baeg, Florida State University
LeCia Barker, University of Texas
Matthew Barr, University of Glasgow
Jo Bates, University of Sheffield
Peter A. Bath, University of Sheffield
Bradley Wade Bishop, University of Kentucky
Pia Borlund, Royal School of Library and Information Science
Fatima Boujarwah, Kuwait University
Jo Ann M. Brooks, Syracuse University
Peter Brusilovsky, University of Pittsburgh
Brian Butler, University of Maryland, College Park
Donald Case, University of Kentucky
Tiffany Chao, University of Illinois
Daegon Cho, Carnegie Mellon University
Namjoo Choi, University of Kentucky
Nicolas Christin, Carnegie Mellon University
Eric Cook, University College Dublin
Andrew Cox, University of Sheffield
Mike Crandall, University of Washington
Morgan Daniels, University of Michigan
Joy Davidson, Digital Curation Centre
Andy Dawson, University College London
Robert Deng, Singapore Management University
Robert Deng, University of Pittsburgh
Lee Komito, University College Dublin
Ellen Detlefsen, University of Pittsburgh
Jana Diesner, University of Illinois
Yunfei Du, University of North Texas
Alison Langmead, University of Pittsburgh
Michael Khoo, Drexel University
Yunhyong Kim, University of Glasgow
Margaret Kipp, University of Wisconsin-Milwaukee
Joyce M. Latham, University of Wisconsin-Milwaukee

* From the University of North Texas
<table>
<thead>
<tr>
<th>Reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diane Leiva, Florida State University</td>
</tr>
<tr>
<td>Päivi Majaranta, University of Tampere</td>
</tr>
<tr>
<td>Bonnie Mak, University of Illinois</td>
</tr>
<tr>
<td>Kathryn Masten-Cain, University of North Texas</td>
</tr>
<tr>
<td>Kate McDowell, University of Illinois</td>
</tr>
<tr>
<td>Andrew McHugh, University of Glasgow</td>
</tr>
<tr>
<td>Pam McKinney, University of Sheffield</td>
</tr>
<tr>
<td>Erik Mitchell, University of Maryland, College Park</td>
</tr>
<tr>
<td>William E. Moen, University of North Texas</td>
</tr>
<tr>
<td>Scott Nicholson, Syracuse University</td>
</tr>
<tr>
<td>Marlies Olensky, Humboldt-Universität zu Berlin</td>
</tr>
<tr>
<td>Sameer Patil, Helsinki Institute for Information Technology</td>
</tr>
<tr>
<td>Celeste Lyn Paul, University of Maryland, Baltimore County</td>
</tr>
<tr>
<td>Wilhelm Peekhaus, University of Wisconsin-Milwaukee</td>
</tr>
<tr>
<td>Vivien Petras, Humboldt-Universität zu Berlin</td>
</tr>
<tr>
<td>Arcot Rajasekar, University of North Carolina at Chapel Hill</td>
</tr>
<tr>
<td>Lionel Robert, University of Michigan</td>
</tr>
<tr>
<td>Howard Rodriguez-Mori, Florida State University</td>
</tr>
<tr>
<td>Corinne Rogers, University of British Columbia</td>
</tr>
<tr>
<td>David Saab, Pennsylvania State University</td>
</tr>
<tr>
<td>Reijo Savolainen, University of Tampere</td>
</tr>
<tr>
<td>Steven B. Sawyer, Syracuse University</td>
</tr>
<tr>
<td>Walt Scacchi, University of California, Irvine</td>
</tr>
<tr>
<td>Barbara Schultz-Jones, University of North Texas</td>
</tr>
<tr>
<td>Michael Steven Seadle, Humboldt-Universität zu Berlin</td>
</tr>
<tr>
<td>Barbara Sen, University of Sheffield</td>
</tr>
<tr>
<td>Elizabeth Shaffer, University of British Columbia</td>
</tr>
<tr>
<td>Kalpana Shankar, University College Dublin</td>
</tr>
<tr>
<td>Katie Shilton, University of Maryland, College Park</td>
</tr>
<tr>
<td>Mette Skov, Royal School of Library and Information Science</td>
</tr>
<tr>
<td>Richard Smiraglia, University of Wisconsin-Milwaukee</td>
</tr>
<tr>
<td>Daniella Smith, University of North Texas</td>
</tr>
<tr>
<td>Eero Sormunen, University of Tampere</td>
</tr>
<tr>
<td>Michael Spring, University of Pittsburgh</td>
</tr>
<tr>
<td>Taverekere Srikantaiah, University of Maryland, College Park</td>
</tr>
<tr>
<td>Beth L. St. Jean, University of Maryland, College Park</td>
</tr>
<tr>
<td>Jeff Stanton, Syracuse University</td>
</tr>
<tr>
<td>Juliane Stiller, Humboldt-Universität zu Berlin</td>
</tr>
<tr>
<td>Mega Subramanian, University of Maryland, College Park</td>
</tr>
<tr>
<td>Melissa Terras, University College London</td>
</tr>
<tr>
<td>Martin Thellefsen, Royal School of Library and Information Science</td>
</tr>
<tr>
<td>Clare Thornley, University College Dublin</td>
</tr>
<tr>
<td>Ryan James Turner, Carnegie Mellon University &amp; Instituto Superior Técnico</td>
</tr>
<tr>
<td>Michael Twidale, University of Illinois</td>
</tr>
<tr>
<td>Robert Villa, University of Sheffield</td>
</tr>
<tr>
<td>Jarmo Välimäki, University of Tampere</td>
</tr>
<tr>
<td>Greg Wadley, University of Melbourne</td>
</tr>
<tr>
<td>Jieyu Wang, University of Maryland, Baltimore County</td>
</tr>
<tr>
<td>Yang Wang, Syracuse University</td>
</tr>
<tr>
<td>Kathleen Weaver, University of Maryland, Baltimore County</td>
</tr>
<tr>
<td>Rosina Weber, Drexel University</td>
</tr>
<tr>
<td>Lynn Westbrook, University of Texas</td>
</tr>
<tr>
<td>Kevin White, University of Maryland, Baltimore County</td>
</tr>
<tr>
<td>Peter Willett, University of Sheffield</td>
</tr>
<tr>
<td>Susan Wilson, University of Maryland, College Park</td>
</tr>
<tr>
<td>Dietmar Wolfram, University of Wisconsin-Milwaukee</td>
</tr>
<tr>
<td>Adam Worrall, Florida State University</td>
</tr>
<tr>
<td>Judith Wusteman, University College Dublin</td>
</tr>
<tr>
<td>Iris Xie, University of Wisconsin-Milwaukee</td>
</tr>
<tr>
<td>Alyson Young, University of Maryland, Baltimore County</td>
</tr>
<tr>
<td>Oksana L. Zavalina, University of North Texas</td>
</tr>
<tr>
<td>Luke Zhang, Pennsylvania State University</td>
</tr>
<tr>
<td>Michael Zimmer, University of Wisconsin-Milwaukee</td>
</tr>
</tbody>
</table>
Preface

The theme of iConference 2013 is Data, Innovation, and Wisdom, an apt set of concepts for the trailblazing ideas and work of the iSchools. The pioneer spirit of iSchools researchers underscores the appropriateness of the conference venue in Fort Worth, Texas. This colorful city, said to be "where the West begins," celebrates the nation's heritage of exploration and expansion into new frontiers. In this vein, we have moved the iSchools forward in publishing with the *iConference 2013 Proceedings*.

This is the first iConference proceedings volume to be published formally as a book with ISBN and ISSN under the name of the iSchools as publisher. ISBN and ISSN are required for major scholarly indexing services that the iSchools may consider in the future.

Last year the iDeans officially agreed that iConference proceedings would be open access, with authors retaining intellectual property rights to their own work. For 2013 we resumed the practice of archiving conference materials in the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS; [https://www.ideals.illinois.edu/handle/2142/34699](https://www.ideals.illinois.edu/handle/2142/34699)). We endeavored to improve access to objects in the repository several ways.

- This is the first time persistent DOI have been assigned to the proceedings, with the aim of facilitating online discovery and retrieval. Both the compiled proceedings book and individual contributions are downloadable and have their own DOI.
- Through CrossRef (http://www.crossref.org/), the DOI registration agency, citation links in the text are incorporated in a global citation linking network encompassing millions of resources. Earlier proceedings can be added to the network retroactively.
- We created more detailed metadata for each object in IDEALS than has been done in the past and imposed name authority control.
- We inserted an author copyright statement in each individual contribution.

This is the first time a complete e-book format has been used for the proceedings, and we endeavored to design it for optimal navigability and readability.

- This publication is downloadable in formats for reading on mobile devices.
- The page design is a simplified single-column style for reading on mobile devices.
- Digital pages can be annotated by users of mobile devices with appropriate apps.
- Table of Contents items are internally linked to each contribution.
- Citations are linked to external resources, and thus are part of the citation linking network.
- Authors’ email addresses are hot-linked to facilitate communication.
- The full text is searchable.

*iConference 2013 Proceedings* contributions are organized into seven categories, including work by new and student researchers and descriptions of conference events. The book contains 210 contributions totaling 1,133 pages. These, along with related materials uploaded to IDEALS later, comprise the largest set of iConference materials to date.

We are pleased to make this publication publicly available and look forward to seeing the iSchools publishing model evolve in coming years.

Linda Schamber, Editor
Proceedings Co-Chair
Contents

- Papers
- Notes
- Posters
- Workshops
- Alternative Events
- Research Paper Development Roundtable
- Social Media Expo

Papers

Developing Computational Supports for Frame Reflection
Eric P. S. Baumer, Francesca Polletta, Nicole Pierski, Chris Celaya, Karen Rosenblatt, Geri K. Gay 1

An Examination of How the FCC Uses “Voluntary Commitments” from Merging Telecommunications Companies to Advance Policy Goals
Gwen Shaffer, Scott Jordan 15

Buying numbers: An Empirical Analysis of the IPv4 Number Market
Milton L. Mueller, Brenden Kuerbis 27

An Exploratory Study of Users’ Preferences and Use of Sources for Seeking Health Information
Yan Zhang 38

Gender as a Predictive Factor for Tasks Completed Using Smartphones
Mark Evans, Susan Hopper, Greg Jones, Gerald Knezek 50

If We Build it, Will They Come? Designing a Community-Based Online Site for Parents
Sarita Yardi Schoenebeck, Amy Bruckman 54

Jobs and Family Relations: Use of Computers and Mobile Phones Among Hispanic Day Laborers in Seattle
Luis Fernando Baron, Moriah Neils, Ricardo Gomez 65

Connecting Government, Libraries, and Communities: Information Behavior Theory and Information Intermediaries in the Design of the LibEGov Tool
Paul T. Jaeger, Ursula Gorham, John Carlo Bertot, Natalie Greene Taylor, Elizabeth Larson, Ruth Lincoln, Jonathan Lazar, Brian Wentz 74

A Framework for Comparing Privacy States
Alan Rubel, Ryan Biava 85

Service Learning in Action: Integrating Reflection to Deepen the Educational Experience
Benjamin K. Baron, Barbara Schultz-Jones 91

Understanding Large Scale Online Environments with Qualitative Methods
Dana Rotman, Yurong He, Jennifer Preece, Allison Druin 100

An Analysis of Cooking Queries: Implications for Supporting Leisure Cooking
Sally Jo Cunningham, David Bainbridge 112

Empirical Analysis of Data Breach Litigation
Sasha Romanosky, David Hoffman, Alessandro Acquisti 124

Hand in Glove, or Square Peg, Round Hole: An Exploration of Social Network Analysis Applied to Online Community Trace Data
Kimberly Glasgow 138

From Push Brooms to Prayer Books: Social Media and Social Networks During the London Riots
Kimberly Glasgow, Clayton Fink 155

Collaboration in Isolation: Bridging Social and Geographical Boundaries in Two Rural Technology Firms
Sean P. Goggins 170

Group Polarization in Virtual Communities: The Case of Stock Message Boards
Hsuanwei Michelle Chen 185

Micro-serendipity: Meaningful Coincidences in Everyday Life Shared on Twitter
Toine Bogers, Lennart Björneborn 196

Gradation: A Pay-as-You-Go Style Hybrid Query Language for Structured and Text Data
Yui Yasunaga, Atsuyuki Morishima, Hiroki Sodeyama, Masateru Tadaishi, Shigeo Sugimoto 209
A Dyadic Approach to Information Mediation at Work: Examining Credibility and Value Perceptions  
Ji Yeon Yang, Soo Young Rieh  
223

Robotics Narratives and Networks: Conceptual Foundations for a Non-linear Digital Archive  
Staša Miljević, Selma Šabanović  
235

Towards a Data Literate Citizenry  
Michael B. Twidale, Catherine Blake, Jon Gant  
247

Cooperative Visualization: A Design Case  
Nathan Prestopnik  
258

Scientific Referential Metadata Creation with Information Retrieval and Labeled Topic Modeling  
Han Jia, Xiaozhong Liu  
274

Understanding the EMR Error Control Practices among Gynecologic Physicians  
Ritu Khare, Yuan An, Sandra Wolf, Paul Nyirjesy, Longjian Liu, Edgar Chou  
289

Effects of User Identity Information On Key Answer Outcomes in Social Q&A  
Erik Choi, Craig R. Scott, Chirag Shah  
302

TagPad: Supporting Field Interviews and Analysis  
Nis Bornoe, Louise Barkhuus  
316

A Study on Q&A Services Between Community-based Question Answering and Collaborative Digital Reference in Two Languages  
Dan Wu, Daqing He  
326

VizDeck: Streamlining Exploratory Visual Analytics of Scientific Data  
Daniel Perry, Bill Howe, Alicia M. F. Key, Cecilia Aragon  
338

Citizens’ Use of Twitter in Political Information Sharing in South Korea  
Jisue Lee, Hohyon Ryu, Lorri Mon, Sung Jae Park  
351

When “Organizing Becomes an Extension of Your Brain”: DJs, Music Libraries and Information Practices  
Jessica Lingel  
366

Searching for Theory in Metadata  
Ben Li  
377

Playfully Serious Information for Serious Play: The Integration of Community Values in an Information Resource  
Julia Bullard  
389

“The Public is Part of the Audience”: Information Access for Citizens, Democratic Accountability, and Climate Change Science  
Shannon M. Ottmann  
398

A Machine Learning-Based Approach to Predicting Success of Questions on Social Question-Answering  
Erik Choi, Vanessa Kitzie, Chirag Shah  
409

Retrocomputing as Preservation and Remix  
Yuri Takhteyev, Quinn DuPont  
422

Notes

Chatting with Friends Online While Watching a Video: What and How Much Information Is Retained?  
Lin Lin, Chris Bigenho  
433

Information Recording in Archaeological Practice: A Socio-Technical Perspective  
Emad Khazraee  
437

Finding and Figuring Flow: Notes Toward Multidimensional Poetry Visualization  
Katharine Coles, Julie Gonnering Lein  
444

Disparities in Distribution of Information Resources in the City of Philadelphia: Spatial Analysis of Socio-demographic and Information Resource Variables  
Emad Khazraee, Adam Townes, Anthony Grubesic  
449

Identifying Crossover Documents in an Interdisciplinary Research Environment  
Rich Gazan  
457
Are You Satisfied Yet? Shared Leadership, Trust and Individual Satisfaction in Virtual Teams
Lionel Robert, Sangseok You

Victoria L. Lemieux

Using Vignettes to Study Conflict Management Practices in Enterprise Architecture Initiatives
Guangxuan Zhang, Sandeep Purao

Information Flows in Events of Political Unrest
Karine Nahon, Jeff Hemsley, Robert M. Mason, Shawn Walker, Josef Eckert

E-books in Academia: Expectations and Challenges
Irene Lopatovska, M. Cristina Pattuelli, Leanora Lange, Victoria Ludas Orlofsky

What Influences the Decision to Participate in Audience-bounded Online Communities?
Claudia López, Rosta Farzan, Shaghayegh (Sherry) Sahebi, Peter Brusilovsky

MASHing Metadata: Legacy Issues in OAI Harvesting From Three Digital Libraries
Michael Khoo, Doug Tudhope, Ceri Binding, Hilary Jones, Ivan Orrego, Jae-wook Ahn

Certain Characteristics of iSchools v. Other LIS Programs
Robert Wedgeworth

“Sometimes I believe as many as six impossible things before breakfast”: Outsider Naïveté as an Asset in the Initiation of Massive-scale Book Provision Efforts
Elisabeth A. Jones

Augmenting Optical Character Recognition (OCR) for Improved Digitization: Strategies to Access Scientific Data in Natural History Collections
Deborah L. Paul, P. Bryan Heidorn

Memory Cues, Recall Strategies, and Alzheimer’s Disease
Lynne C. Howarth, Erica Hendry

Do Members Converge to Similar Reasoning Styles in Teamwork? A Study of Shared Rationales in Small Team Activities
Lu Xiao

Educational Dialogue and Technological Utterances: A Phenomenological Approach to Developing a Theory of Evocative Design Practice
Michael Marcinkowski

Comparative Analysis of User Searching in Domain-Specific and Domain-Independent Digital Libraries
Oksana L. Zavalina, Elena V. Vassilieva

Utilizing Public Displays to Enable Location-Centric Information Sharing
Nicholas Wittison, Christopher Lueg

Context and Collection: A Research Agenda for Small Data
Amelia Abreu, Amelia Acker

“I’ll Be Watching You”: What do Canadians Know About Video Surveillance and Privacy?
Brenda McPhail, Joseph Ferenbok, Roxanna Dehghan, Andrew Clement

Towards a Method of Documentary Practices for Personal Health Information Management
Matt Willis, Carsten Østerlund

Intelligent Urbanism: Convivial Living in Smart Cities
Stephanie Santoso, Andreas Kuehn

Information Management and Communication for Dementia: Preliminary Research from China
Claire L. Barco, Koji Yatani, Yuanye Ma, Candra K. Gill, Joyojeet Pal

Exploring Hybrid-Economic Communities and the Technology-Mediated Identities Performed There
Warren Allen

It Takes a Jerk to Make a Conversation into an Archive
Julia Bullard

iPad as PIM Device: Input, Interrelation and Indexes
Paris Buttfield-Addison, Christopher Lueg, Leonie Ellis

Learning with Social Media: Measurement Tools for Understanding Information Behavior in Technology Pervasive Environments of the 21st Century
Leila A. Mills, Gerald A. Knezek, Jenny S. Wakefield
Visualizing Overlapping Latent Communities Using POI-Based Visualizations
Patrick M. Dudas, Jae-wook Ahn, Martijn de Jongh, Peter Brusilovsky 601

The Atlas Project—Comparing Time and Place
Tim Nugent, Christopher Lueg, Andrew Bennett 611

Regulating Anti-Social Behavior on the Internet: The Example of League of Legends
Yubo Kou, Bonnie Nardi 616

Cooperative, Dynamic Twitter Parsing and Visualization for Dark Network Analysis
Patrick M. Dudas 623

Re-Inventing the Wheel: The Re-Creation of Documents in a Bumble-Bee Organization
Ingbert Schmidt 633

Hanna Maurin Söderholm, Monica Lassi, Gustaf Nelhans, Jan Nolin 638

Is the Article Crucial to My Research? Evaluating Task-Oriented Impacts of Scientific Articles in Information Seeking
Jiepu Jiang, Daqing He, Shuguang Han, Wei Jeng 645

---

Posters

Instructional Strategies for Autistic Adults Learning Video Game Design
Demetria Ennis-Cole, Princess Cullum, Gary Moore 650

Fearless Cards: Computer Training for Extremely Marginalized Populations
Ricardo Gomez, Ivette Bayo, Philip Reed, Cherry Wang 653

iDiscover: Inspiring Youth to Pursue STEM
Daniel Knopp, Olivia Green, Clea Counts 654

DataUp: Enabling Data Stewardship for Researchers
Carly Strasser 657

The “Forever Problem”: Nuclear Waste as Information
Chris Heaney 659

When Thumbnails Are and Are not Enough: Factors Behind Users’ Requests for Information from a Video Digital Library
Dan Albertson 662

Meta-Scraping: Two Technological Approaches to Support Meta-analyses
Kim Nimon, Cornelia Caragea, Frederick L. Oswald 665

International Comparative Study Analyzing Knowledge Organization Research Topics in Four Countries – Brazil, South Korea, Spain and United States
Jihee Beak, Jeanette Glover, Daniel Martinez-Ávila, Suellen Oliveira Milani 668

Two Paths to Motivation through Game Design Elements: Reward-Based Gamification and Meaningful Gamification
Scott Nicholson 671

Smart TV as a Social Platform
Dong-Hee Shin, Taeyang Kim 673

Cardiac Rehabilitation Referral: An Information Gap
Claudia J. Gollop, Annie T. Chen, Wan-Ching Wu 682

Lights, Cameras, Action: Using Video to Create Instructor Presence in an Online Information Studies Course
Sharon Stoerger 685

Proposed Facets of a Serendipitous Digital Environment
Lori McCay-Peet, Elaine G. Toms 688

Large-Scale Digital Library User Searching: What Role Does Domain Play?
Oksana L. Zavalina, Elena V. Vassilieva 692

The Neighborhood Book Exchange: Community Catalyst or Media Hype?
Kathleen Gollner, Tenny Webster, Lisa Nathan 697
Placing the Walls: 3-D Modeling of Chicana/o Murals in LA
   David Kim
Measuring Serendipity in the Lab: The Effects of Priming and Monitoring
   Toine Bogers, Rune Rosenborg Rasmussen, Louis Sebastian Bo Jensen
The Role of Digital Libraries as Boundary Objects Within and Across Communities
   Adam Worrall
Automatic Detection of Search Tactic in Individual Information Seeking: A Hidden Markov Model Approach
   Shuguang Han, Zhen Yue, Daqing He
An Insight Into Vector Space Modeling and Language Modeling
   Kun Lu
Watching for Whales: Emerging Configurations of Scholarly Communication
   John S. Seberger, Cory P. Knobel
Network Maps and Congressional Frames: Analyzing Bill Titles as a Field of Conflict
   Ryan Biava
Real Talk: A Toolkit for Community Engagement, Transparency, and Mobile Governance
   Hale Thompson, Jes A. Koepfler, Katherine Sydenham, Anthony Hoffman
Information Occupation: Using Information Science to Explore Occupy Wall Street
   Oliver L. Haimson, Josh Cartagena
A Practice Perspective on Websites for the Sharing Economy
   Gabriel H. Mugar
Cultural Context of Social Media Use: How Do Iranians Adapt and Use Facebook?
   Emad Khazraee
Characterizing the Scholar H-index Via Full-text Citation Analysis
   Star X. Zhao, Xiaozhong Liu, Fred Y. Ye
Developing Community-Centered Programs for Rural Information Professionals
   Yunfei Du
Exploring Cyberbullying Through Visual Narratives
   Leanne Bowler, Eleanor Mattern, Cory Knobel, Patrick Keilty
Paradigm Shift of LIS Education in the North East Region India
   Bobby Phuritsabam, Th. Purnima Devi, Memori Sagolsem
Using Digital Book Metrics for Navigation and Browsing
   Michael Huggett, Edie Rasmussen
The Usability Study on the Multicultural Children’s Book Project of the National Library for Children
   and Young Adults (NLCY) in Korea
   Wooseob Jeong, Renee Kapusniak, Hyejung Han
Examining Transdisciplinarity in iSchools Dissertations (2010-2011)
   Kathryn Masten
Project VIEWS2—Phase 1: Innovation in Early Literacy
   Janet L. Capps, Ivette Bayo, Katie Campana, Erika N. Feldman, Eliza Dresang, Kathleen Burnett
Scientific Metadata Quality Enhancement for Scholarly Publications
   Chun Guo, Jinsong Zhang, Xiaozhong Liu
Teaching Tools, Applications, and Infrastructure for Digital Curation Through the Use of a Virtual Lab
   Joseph Helsing, Paulette Lewis, Edward Warga
A Prototype System for Heterogeneous Data Management and Medical Devices Integration in
   Trauma Resuscitation
   Zhan Zhang, Aleksandra Sarcevic, Yuan An
Integrating Data Curation Concepts Throughout the Project Lifecycle: A WILIS Case Study
   Cheryl A. Thompson, Joanne Gard Marshall, Jennifer Craft Morgan, Susan Rathbun-Grubb, Amber Wells
Interacting with Health Information for Self-care: A Pilot Study Exploring Undergraduates’
   Health Information Literacy
   Jinxuan Ma, Don Latham
A Comparative Study of Data Reuse Among Quantitative Social Scientists and Archaeologists
   Ixchel M. Faniel, Julianna Barrera-Gomez, Adam Kriesberg, Elizabeth Yakel
Authenticity as a Social Contract—We Are Our Records
Corinne Rogers, Joe Tennis

Randomized Sampling: An Approach to Extraction of Metadata Records
Olajumoke Azogu, Jiangping Chen

Organizational Perspectives of Open Innovation in Government
Claudia A. Louis

How Information Science Professionals Add Value in a Scientific Research Center
Christopher Eaker, Andrea Thomer, Erica Johns, Kayla Siddell

Using Design Thinking to Empower Ethnic Minority Immigrant Youth in their Roles as Information and Technology Mediaries
Karen E. Fisher, Phil Fawcett

Old Maps and Open Data Networks
Bernhard Haslhofer, Keith Newman, Amanda Stefanik, Werner Robitza, Carl Lagoze

Literature Review: How Content Analysis Is Used to Study Human Rights Violations in Political Science?
Tatiana Vashchilko, Lu Xiao

Research Specialties as Emergent Phenomena: Connecting Emergence Theory and Scientometrics
T. Clay Templeton, Kenneth R. Fleischmann

Digital Curation Tools: Metadata Enhancement with Selenium IDE
Daniel Gelaw Alemneh, Andrew James Weidner

Looking @ Position: Examining the Location of Twitter Handles in a Political Event
Christopher Mascaro, Nora McDonald, Alan Black, Sean Goggins

Reading Practices and Intellectual Freedom Research
Emily J.M. Knox

Open Access Publishing Fees: Responses, Strategies, and Emerging Best Practices
Laura Waugh

Information Behavior of Rural Dementia Caregivers Over the Course of Counseling Interventions
Michelle M. Kazmer, Robert L. Glueckauf, Jinxuan Ma, Gabriel Schettini, Michelle Silva

Can’t See the Forest for the Trees? A Citation Recommendation System
Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra, C. Lee Giles

“I have to read the signs like you reads a book”: Examining Experiential Information through Non-fiction Accounts of Guided Trips into the Interior of Algonquin Provincial Park Ontario
Angela Pollak

Enhancing Content Visibility in Institutional Repositories: Overview of Factors that Affect Digital Resources Discoverability
Ahmet Meti Tmava, Daniel Gelaw Alemneh

A Participatory Framework for Evaluation Design
Diana S. Kusunoki, Aleksandra Sarcevic

Practical Approaches and Proposed Strategies for Measuring Selected Aspects of Community-Based Broadband Deployment and Use
Lisandra R. Carmichael, Charles R. McClure, Lauren M. Mandel, Marcia Mardis

The Lifecycle Management of ETDs Project: Multi Stakeholders National Partnership
Shannon Stark, Daniel G. Alemneh, Stephen Eisenhauer

Implementing an Open Source EMR in a Nursing Informatics Course
Nina Multak, Emad Khazraee, Michelle Rogers, Prudence Dalrymple

Toward a Mesoscopic Analysis of the Temporal Evolution of Scientific Collaboration Networks
Syed Ishtiaque Ahmed, Scott Allen Cambo, Carl Lagoze, Theresa Velden

Label Annotation through Biodiversity Enhanced Learning
P. Bryan Heidorn, Qianjin Zhang

Surveillance of Audience Labor Using New Media: Three Innovations of Television Broadcast Networks
Brian L. Griffin

Public Opinion Aggregation by Annotation and Tagging of Online News Stories
Pranay Sethi

801
804
808
812
816
819
823
827
831
836
840
842
844
849
852
855
860
865
869
874
878
882
885
891
Personal Information, Personal Property
Laura Lenhart
895

Information Systems and Smart Grid: New Directions for the IS Community
Jason Dedrick, You Zheng
897

Ebooks and Cross Generational Perceived Privacy Issues
Michael Zimmer, Jennifer Thiele, Renee Kapusniak
900

Using Machine Learning Models To Interpret Disciplinary Styles of Metadiscourse in Dissertation Abstracts
Bradford Demarest, Cassidy Sugimoto
901

Extending the VIVO Ontology to iSchools: Enabling Networking of Information Scientists
Ying Ding, Erjia Yan, Ali Ghazinejad, Han Jia
905

The It Girls Overnight Retreat: An iSchool Recruitment Model for Prospective Female Undergraduate Students
Dori Farah, Julie Walas
909

The Globalization of the iSchools Movement
Laurie J. Bonnici, Heidi Julien, Kathleen Burnett
912

Community Informatics Studio: A Model of Information Scholarship in Action
Martin Wolske, Colin Rhinesmith, Jennie Archer, Emily Bayci, Ryne Leuzinger, Lucas McKeever
916

Nonparametric Estimation of Search Query Patterns
Soo hyung Joo, Dietmar Wolfram, Suyong Song
919

HELPP Zone: Towards Protecting College Students from Dating Violence
Amirreza Masoumzadeh, Lei Jin, James Joshi, Rose Constantino
925

Academic Genealogy as an Indicator of Interdisciplinarity: A Preliminary Examination of Sociology Doctoral Dissertations
Cha oqun Ni, Cassidy R. Sugimoto
929

Dispatches, Digests and Doodles: Exploring the Significant Properties of Field Notebooks
Andrea K. Thomer, Nicholas M. Weber
937

Identifying Claims in Social Science Literature
Shameem Ahmed, Catherine Blake, Kate Williams, Noah Lenstra, Qiyuan Liu
942

Participation in the City: A Typology of Open Government Data Use
Morgan Currie
947

Lead, Lag or Get Out of the Index: Exploring Macro-economic Indicators of Data Use
Nicholas M. Weber, Andrea K. Thomer
949

A Model for Assessing the Quality of Gene Ontology
Shuheng Wu
953

Improving the Character of Optical Character Recognition (OCR): iDigBio Augmenting OCR Working Group Seeks Collaborators and Strategies to Improve OCR Output and Parsing of OCR Output for Faster, More Efficient, Cheaper Natural History Collections Specimen Label Digitization
957

Using Etherpads as Platforms for Collaborative Learning in a Distance Education LIS Course
Bob Pym m, Lyn Hay
965

SoMe Tools for Social Media Research
Shawn Walker, Jeff Hemsley, Josef Eckert, Robert Mason, Karine Nahon
971

Remediating Tinker Bell: Childhood Commodification and the Transmedia Narrative
Eric M. Meyers, Lindsey Krabbenhoft, Julia P. McKnight
972

Trust and Community: Continued Engagement in Second Life
Peyina Lin, Natascha Karlova, John Marino, Michael B. Eisenberg
976

New Models for Supporting Digital Scholarship: UNT’s DiSCo
Anjum Najmi, Spencer D. C. Kerals
981

Spontaneous Reactions to an Anti-piracy Initiative: A Youtube Clip Micro Analysis
Gustaf Nelhans, Hanna Maurin Söderholm, Jan Nolin, Mathias Klang, Monica Lassi
985
Workshops

**Sociotechnical Systems Research: Defining, Converging, and Researching as a Community**
Rebecca Reynolds, Warren Allen, Ingrid Erickson, Shuyuan Mary Ho, James Howison, Ingbert Schmidt

**Educating E-Discovery Litigation Support Professionals**
Jason R. Baron, Douglas W. Oard

**Digital Youth Workshop: Calling All Designers, Researchers, and Policy Makers**
Karen E. Fisher, Eliza T. Dresang, Katie Davis, Allison Druin, Sarita Yardi

**An Introduction to Policy Informatics**
Erik Johnston, Kevin DeSouza, Tanya M. Kelley

**Big Social Data for Social and Information Scientists**
Sean Goggins, Christopher Mascaro, Nora McDonald, Alan Black, Guiseppe Valetto

**Data Management Planning: Where We’ve Been and Where We are Going Next**
Carly Strasser, Martin Donnelly, Sherry Lake

**Filling the Workforce Gap in Data Science and Data Analytics**
Catherine Blake, Jeffrey M. Stanton, AnnaLee Saxenian

**Computational Scientometrics: Theory and Applications**
Cornelia Caragea, C. Lee Giles, Vette I. Torvik, Lior Rokach

**Help iDigBio Reveal Hidden Data: iDigBio Augmenting OCR Working Group Needs You**
Deborah Paul, P. Bryan Heidorn, Jason Best, Edward Gilbert, Amanda Neill, John Mignault

**From iSchools to US Ignite: Strategies for Fostering Inclusive Gigabit Networks**
Jon P. Gant

**iSchools Building on the Strengths Found in the Convergence of Librarianship, Archival, and Museum Studies to Improve the Education of Managing Digital Collections**
Anna Maria Tammaro, Vittore Casarosa, Seamus Ross, Heather Moulaison, Terry Weech, Frederick Lugya

**Special Workshop on Information Privacy**
Alessandro Acquisti, Deirdre Kathleen Mulligan

Alternative Events

**Educating Information Professionals for 2050**
Gary Marchionini, Mike Eisenberg, Liz Liddy, Cathy Marshall

**Methods and Technologies to Promote Information-Centered Knowledge Construction**
J. Michael Spector, Dirk Ifenthaler, Gerald Knezev, Tandra Tyler-Wood, ChanMin Kim

**The Problem of Data: iSchools and the New Data Professional**
Spencer D. C. Keralis, Shannon Stark

**Calling All Computer Scientists and Social Scientists: Establishing a Research Agenda for Computational Social Science**
Kenneth R. Fleischmann, Sean P. Goggins, James Howison, Matthew Lease, Douglas W. Oard
Experience Science in Informal Learning Environments: Tales from the Field
Mega Subramaniam, June Ahn, Nicole D. Alemanne, Mike Eisenberg, Sean Fullerton, Michelle M. Kazmer, Marcia Mardis

Whither the Child? The iSchool Approach to Research Relating to Children and Adolescents
Leanne Bowler, Kate McDowell, Denise Agosto

Values Interventions: Ethics Scholarship in Action
Katie Shilton, Michael Zimmer, Finn Brunton, Deirdre Mulligan

Plan Play Pressure Pause: Engaging Creative Information Practices
Theresa Anderson, Leanne Bowler, Lisa Nathan, Eileen Trauth

Theory and the Social Nature of Information
Adam Worrall, Jes A. Koepfler, Jessica Vitak, Nicole D. Alemanne, Jaime Snyder, Howard Rosenbaum

“It’s a Series of Tubes”: Exploring Net Neutrality Policy through Critical Making
Jean-François Blanchette, Matt Ratto

Design Jams in iSchools: Approaches, Challenges and Examples
Michael B. Twidale; Ingbert Schmidt; Jeff Ginger; Peter Organisciak; Piotr Adamczyk; Christopher Lueg

Igniting Talk on Digital Literacy
Ingrid Erickson, Eric Meyers, Ruth Small

Tech Learning Potluck
Michael B. Twidale, Sally Jo Cunningham, James Howison, Peter Organisciak, Jeff Ginger, Ingbert Schmidt, Brittany Smith, Karen Baker, Kate Williams

Collaboration in Action: Enabling Innovative Scholarship with Social and Crowdsourcing Services
Chirag Shah, Daqing He, Marie L. Radford, Jung Sun Oh, Lynn Silipigni Connaway

Methods, Practice, Theory: Researching Human Information Behavior in the Context of Migration
Jessica Lingel, Lisa Quirke, Nadia Caidi, Karen Fisher, Ellen Pozzi

Help iDigBio Reveal Hidden Data: iDigBio Augmenting OCR Working Group Needs You -- Part II
Deborah Paul, P. Bryan Heidorn, Jason Best, Edward Gilbert, Amanda Neill, John Mignault

Research Paper Development Roundtable

Social Values in a Future Internet: Analyzing the Named Data Networking Protocols
Katie Shilton

Sociomateriality in the Sand Box: Visually Oriented Documenting Practices of Young Entrepreneurs
Jamie Snyder, Carsten Østerlund

Highly Dedicated Enthusiasts: Online Museum Visitors Doing Everyday Life Information Seeking
Mette Skov

Conceptualizing the Information Field: A Multidimensional Approach Based on Interdisciplinary Theory
Dorte Madsen

Patient-Provider Communication and Information Access via Online Patient Portal: Understanding User Behaviors to Improve Healthcare and Outcomes
Xiaomu Zhou, Si Sun

Sociomateriality of Information Standards in Infrastructure: Structure, Organization, and Process Constraints
Jo Ann Brooks, David Mann

“Citizen Science Online Gaming: Work or Play?” or “Some of the Values in the Design of Web-based Citizen Science Games: The Case of Phylo”
Brian Beaton

1037
1039
1041
1044
1047
1051
1053
1056
1058
1061
1063
1066
1069
1070
1071
1072
1073
1074
1075
Social Media Expo

Learning Focused Search in an Online Social Network Community
Brittany N. Smith, Fredrick K. Lugya, Craig Evans 1076

K-Pop Live: Social Networking & Language Learning Platform
Thomas Chua, Aloysius Lau, Chin Leng Ong, Kian Ming Png, Houston Toh, Kyong Jin Shim, Feida Zhu, Ee-Peng Lim 1081

Cairn: Using Digital Markers to Bridge Learning Communities
Amanda Menking, Ivette Bayo, Sean Fullerton 1086

Learning with Polyphony: AmpDamp
Amelia Acker, Melissa Chalmers, Matt Burton, Grant Wythoff, Thomas Lodato 1090

Towards a Methodology of Virtually Augmenting a Knowledge Sharing Community of Practice: A Case Study of the Local Food System of Denton, Texas
Andrew Miller, Laura Pasquini, Leila Mills, Mark Evans 1095

Scientific INQuiry (SINQ): Social Media for Everyday Science Learning
Michael Gubbels, Jason Yip, Jinyoung Kim, June Ahn 1102

iMapLibraries: Mapping Opportunities for Lifelong Learning
Laura Spears, Nathaniel Ramos Jr., Jisue Lee 1106

#ElectionClass: Encouraging Strategic Thinking About Politics Through Social Media
Chelsea Danielle Orcutt, Jared Scott Mandel 1110

Analyzing How Users Utilize riff for Collaborative Searching and Sharing Contents for Social Learning in So.cl
Erik Choi, Vanessa Kitzie, Chathra Hendahewa 1114

Collaborative Timeline: Toward the Next Generation of History Learning
Yanqi Wu, Jin Yan, Yanting Zhang, Xiaohan Yu 1118

UPitt iRiS Lab at iConference 2013 Social Media Expo: Search, Share and Learn the Way You Want
Zhen Yue, Wei Jeng, Jiepu Jiang, Shuguang Han, Daqing He 1125

The Need for Information Gurus in Online Learning Communities
Nick Josten, Jonathan Kift, Allison Trumble, Brigid Winter 1130
Developing Computational Supports for Frame Reflection

Eric P. S. Baumer¹,³, Francesca Polletta², Nicole Pierski², Chris Celaya², Karen Rosenblatt³, Geri K. Gay¹,³

¹Communication, Cornell U
{ericpsb, gkg1} @ cornell.edu

²Sociology, U of CA, Irvine
{polletta, npierksi, ccelaya} @ uci.edu

³Information Science, Cornell U
krr46@cornell.edu

Abstract

As the number and variety of sources for political information increase, it can become difficult to attend to the complexities of political issues. This difficulty lies not only in understanding what is being said, i.e. the content of an issue, but also how it is being said, i.e., the framing of the issue. This paper presents a prototype visualization tool designed to encourage attention to, and critical reflection about, the ways in which a political issue is framed. The tool visually presents linguistic analysis of documents about the issue of cap and trade. Results show that tool use interacted with participants’ prior views in affecting their ability to suggest novel framings of the issue, one potential indicator of frame reflection. Tool use also mediated participants’ exposure to different viewpoints. These findings help provide insights on how the design of tools for civic participation can help promote thoughtful, reflective political engagement.

Keywords: political framing, frame reflection, natural language processing, information visualization, sustainability

Introduction

An increasing amount of political content, information, and discussion occurs online, both in relatively novel media, such as blogs, as well as through online versions of more traditional news sources, such as newspapers. However, much of the political discussion online is highly balkanized (Adamic & Glance, 2005; Hargittai, Gallo, & Kane, 2007), with some work suggesting that many people simply surround themselves with like-minded sources (Garrett, 2009; Munson & Resnick, 2010).

This divisiveness results from more than just differences in interests, priorities, or values. Rather, the seeming irreconcilability of political controversies is often due to differences in how issues are framed. Terms such as “tax relief,” “death panels,” “global warming,” and “racial quotas” have famously rallied citizens around causes that are actually quite complex. Furthermore, the way in which an issue is framed—how a problem is explained, to what other problems it is linked, how potential solutions are evaluated, etc.—has a significant impact on people’s perception of the issue and their prescriptions for action (Hart, 2010; Maibach, Roser-Renouf, & Leiserowitz, 2009; Price, Nir, & Cappella, 2005). A related vein of research on “frame reflection” (Schön & Rein, 1994), critical thinking about how issues are framed, has shown that people are more likely to reach agreement on an issue if they can first come to understand the various frames being applied, both others’ and their own.

The question, then, is how to encourage attention to issues of framing. Aside from the challenge of convincing people to scrutinize their own assumptions, identifying and analyzing frames requires close, detailed reading of texts, analysis and coding by trained researchers, and synthesis of those data through

Acknowledgements: This material is based upon work supported by the NSF under Grant No. IIS-1110932, a Cornell Institute for the Social Sciences Small Grant, and a University of California, Irvine Center for the Study of Democracy Seed Grant. Thanks to our study participants, and to members of the Interaction Design Lab for helpful conversations.


Copyright is held by the authors.
statistical calculations or expert interpretation (Gamson & Modigliani, 1989). These challenges are exacerbated when considering the volume of political content and discussion produced on a daily basis. Novel computational techniques, however, enable analysis of vast quantities of textual data with relative speed and ease, transforming this challenge—an abundance of content—into a valuable resource: a wealth of data. Since framing is often evidenced by linguistic patterns, it may be possible to leverage existing computational linguistic techniques to draw attention to patterns that are indicative of framing.

In this paper, we suggest that such computational text analysis tools may be useful in promoting political frame reflection. We present a prototype information visualization tool that shows associations between groups of words, identified via selectional preference learning (Resnik, 1993; Ritter, Mausam, & Etzioni, 2010). This tool does not identify frames per se; rather it identifies and presents visually linguistic patterns that may be relevant to framing. This system was evaluated via a controlled laboratory study where participants used the system to familiarize themselves with documents discussing the cap and trade approach to carbon emission regulation. Not only is cap and trade a contentious political issue rife with potential framings, but it also pertains to the recent interest in sustainable HCI, particularly with respect to how HCI can be relevant to political aspects of sustainability (DiSalvo, Sengers, & Brynjarsdóttir, 2010). Our results not only indicate that the system played an important role in aspects of frame reflection, but they also point to certain aspects of the system’s design as more or less effective at promoting attention to framing. Our primary contribution is not the specific design presented here, but rather the notion of computational supports for frame reflection, the insights gained about how we might design them, and suggestions for how we might evaluate those designs. These insights not only provide important guidance for system designers building tools to facilitate political discussion, but they also help us better understand the processes of framing and frame reflection.

Related Work

Political Framing

In order to make sense of their interactions with one another, people frame their experiences (Goffman, 1974). Frames help people “locate, perceive, identify, and label,” that is, organize and give meaning to, information about experience in the world. A frame consists of a variety of components, including “keywords, stock phrases, stereotype images” (Entman, 1993, p. 52), “metaphors, exemplars, catchphrases” (Gamson & Modigliani, 1989, p. 3), and other devices that provide an interpretive lens or “package” (Gamson & Modigliani, 1989) through which to perceive and make sense of facts or events. Frames define what counts as a problem, diagnose what is causing those problems, make moral judgments about problems and those involved, and suggest remedies for resolving the problems (Entman, 1993). Put simply, frames are a way of organizing our world.

Framing can significantly impact perceptions of political issues. For example, in debates over nuclear power, different frames—such as nuclear power as economic development, or nuclear power as Frankenstein’s monster—were used in different historical periods to justify decisions and legislation in support of, or against, nuclear power (Gamson & Modigliani, 1989). Another study (Price et al., 2005) found that, in the context of group discussion about gay civil unions, framing interacted with individuals’ previous political ideologies; participants reacted strongly to a frame that conflicted with their view, becoming more extreme in their position. These studies show how framing can have a significant impact on interpretations of, and recommendations for addressing, contentious issues.

The work presented here draws on these insights but takes a slightly different approach. Rather than arguing for the use of one specific frame or another, we explore the design of systems intended to support awareness of, and reflection upon, framing. Schön and Rein (1994) argue that complex policy debates can become intractable when stakeholders approach the situation using different frames. They describe, as one example, a decade-long policy controversy about homelessness in Massachusetts. At the time, three dominant frames for the issue—social welfare, access to the market, and social control—all provided different prescriptions for addressing homelessness. The controversy was only resolved by a reframing that synthesized key elements from each of these frames. Thus, Schön and Rein argue that the process of frame reflection, critical examination of the various frames applied to an issue and consideration of alternative framings, can enable productive discussion and resolution of such complex
debates. This paper, then, argues for the use of computational tools that draw attention to, and encourage reflection on, framing in the context of political issues.

### Selectional Preference Learning

Many indicators of framing noted above, such as “keywords, stock phrases” (Entman, 1993, p. 52), “metaphors, exemplars, [and] catchphrases” (Gamson & Modigliani, 1989, p. 3), may be evidenced by linguistic patterns. Many computational techniques have been developed for identifying such linguistic patterns. For this tool, we used a technique called selectional preference learning (Resnik, 1993). For example, words for liquids or fluids tend to occur in specific grammatical relationships with verbs such as “pour,” “flow,” “freeze,” or “evaporate.” That is, liquids select for these verbs. Calculating selectional preferences of common terms in political corpora, such as “senate” or “tax,” may help draw attention to how those concepts are being framed.

Several different methods for calculating selectional preferences have been developed, e.g., (Resnik, 1993; Ritter et al., 2010). The tool described here uses Ritter et al.’s (2010) approach, which employs probabilistic topic modeling to calculate relatively robust results from even limited data.

### Textual Visualization and Political Analysis

This paper draws on previous work involving visualizing textual patterns or political data. The online tool ManyEyes (Viégas, Wattenberg, van Ham, Kriss, & McKeon, 2007) provides a few different visualizations for text data. Most of these are relatively simple, involving pure word counts or in some cases co-occurrence data. While potentially interesting, they did not seem particularly well-adapted to our purposes here.

With respect to political data, Opinion Space (Faridani, Bitton, Ryokai, & Goldberg, 2010) uses principal components analysis to project opinions expressed by users onto a two-dimensional “space” of opinions, using patterns that emerge in the data rather than traditional ideological axes. ConsiderIt (Kriplean, Morgan, Freelon, Borning, & Bennett, 2012) uses commenting and summarizing to facilitate the process of multiple parties understanding each other during political discussion. While thematically related, the computational techniques involved in these projects differ from the one used here.

The systems most similar to that in this paper involve computational analysis of political text. ManyBills (Assogba, Ros, DiMicco, & McKeon, 2011) uses topic modeling to analyze text of proposed legislation in the US Congress. The system aims to make apparent what bills are actually about as well as portions of the bill that seem unrelated to the overall focus of the legislation. metaviz (Baumer, Sinclair, & Tomlinson, 2010) uses computational metaphor identification to analyze political blogs in order to identify, and draw attention to, conceptual metaphors. metaViz resembles the tool in this paper, in that the goal is promoting critical thinking and reflection. As such, some components are influenced by the design of metaViz. However, metaphors are only one component of framing. The tool presented here goes further to promote more general frame reflection.

### Implementation and Design Rationale

The work on framing summarized above describes how frames are linked to patterns of language. This section describes a prototype visualization meant to draw attention to one such pattern: associations between nouns and verbs.

### Data

As mentioned above, the data analyzed for this paper pertain to cap and trade. We analyzed selectional preferences in 227 documents totaling 310,235 words. To collect the documents, we began by searching Google with such terms as “cap and trade,” “cap and trade benefits,” “cap and trade costs,” “cap and trade summary,” and other similar queries. The search results were then pruned to include only those that focused on cap and trade, as opposed to, e.g., mentioning cap and trade as one of several legislative items before Congress. The resulting documents came from websites for nonprofit environmental organizations, electrical utility and oil industry groups, governmental agencies, consumer
organizations, and other similar entities. Some documents contained overviews of different approaches to global warming, while others were more focused critiques or endorsements of cap and trade legislation. Some were directed to a general readership while others were targeted to narrower political or industry groups. In addition, we used the Proquest Historical Newspapers archive to locate editorials and opinion pieces about cap and trade that ran in national newspapers between January 1, 2009 and July 2011. An effort was made during document collection to represent the variety of arguments made for and against cap and trade programs. We then sorted documents into three categories based on whether they contained solely arguments in favor of cap and trade (Pro), solely arguments against cap and trade (Con), or a mix of arguments for and against (Mixed). We expected that each of these positions would likely frame cap and trade differently, and allowing users to see this distinction might help them interpret the visualized linguistic patterns. Table 1 summarizes each of these three corpora.

Table 1
Size of each corpus in the data set

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Con</td>
<td>120</td>
<td>153,906</td>
</tr>
<tr>
<td>Mixed</td>
<td>69</td>
<td>107,124</td>
</tr>
<tr>
<td>Pro</td>
<td>38</td>
<td>49,205</td>
</tr>
</tbody>
</table>

These documents were all parsed using a typed-dependency parser (De Marneffe, MacCartney, & Manning, 2006), the results of which were used to calculate selectional preferences (Ritter et al., 2010). Each noun then selects for a group of related verbs with a group of related grammatical relations. For example, the noun “offset” selects for the verbs “sell,” “buy,” “give,” “auction,” and others with “offset” as mainly the direct object and sometimes the passive subject. These selectional preferences are computed for each corpus.

Visualization

The visualization then presents the selectional preference results. The visualization tool described here was developed using D3 [http://mbostock.github.com/d3/]. The tool starts by showing nouns that appear in each corpus with greater frequency than in general English, as derived from (Kilgarriff, 1996). Each word is color coded based on the key to indicate the corpus from which it comes (Figure 1, left). The user can either click on a noun to see with what verbs it is associated, or can use the search box (Figure 1.a) to narrow the list.

Figure 1. Prototype visualization for supporting frame reflection by showing selectional preferences of relatively frequent nouns.
Clicking a noun hides the top-level list and shows the verbs for which that noun selects, grouped by similar grammatical relations. Verbs are scaled based on a logarithmic transform of the strength with which the chosen noun selects for them and are arranged in a circle around the noun (Figure 1.b). We paper prototyped and mocked-up several different layouts for this view, including more traditional tabular styles. Such rectilinear layouts often had a stronger implied ordering and importance, e.g., top-to-bottom, while the circle layout provided greater flexibility in reading order and interpretation. The "offset" example described above is shown in Figure 1. Again, verbs are color-coded to indicate the corpus from which they came. The left-hand side key is cropped here but is still shown in the visualization.

Clicking on a verb shows an example of the association, for example, where "offset" occurs with the verb "allow" (Figure 1.c), with the number in parentheses indicating how many other similar examples there are. Clicking the example fragment opens a side bar that shows all of these similar fragments (Figure 1.d). Clicking any of the fragments then opens the original text of the document from which it came, with the fragment highlighted.

Importantly, the meanings of these associative patterns are left open to interpretation. For example, one might see "offset"'s association with "auction," "purchase," and "trade" as indicative of an economic approach that effectively uses a free-market to reduce carbon. Alternatively, one could read those associations as unduly emphasizing the financial aspects at the expense of discussing actual environmental impacts. Such interpretive flexibility, we suggest, helps in supporting frame reflection.

**User Study**

We conducted an evaluation of this tool to understand how aspects of its design might impact frame reflection. Here, we describe the study methods, the coding scheme used to assess aspects of frame reflection, our approach to analyzing the data, and the results of that analysis.

**Methods**

Participants, recruited from the social science subject pools at two US research universities, were asked to take part in a study about reading political content and were compensated with course extra credit. A total of 53 (30 female, 23 male) participants were recruited.

Participants began by completing a pretest about existing political views, as well as previous knowledge of and attitudes about climate change. The survey included a 15-item inventory on attitudes about global warming (Maibach et al., 2009), questions about how informed the participant felt about global warming, some factual questions about the causes and consequences of global warming, questions about the participant’s prior knowledge of cap and trade, a brief summary of how a cap and trade policy would work, and a question asking whether or not the participant would support such a policy.

Participants were randomly split into two conditions. Those in the experimental condition (N=27) used the visualization tool described above. Those in the control condition (N=26) used a simple document browser: the participant would click on a corpus name, click the title of a document, and read the document (Figure 2). Before using each tool, participants watched a short video describing it. Participants were asked to spend 10 to 20 minutes using the tool and were not able to advance to the next portion of the survey until at least 10 minutes had passed.

Afterward, all participants completed an exit survey, which included the same questions from the initial survey asking about their support for cap and trade, a randomly ordered pair of questions asking the participant to state all the arguments s/he could think of for and against cap and trade (Cappella, Price, & Nir, 2002), a randomly ordered pair of questions asking the participant to name the groups of people who would be most positively affected and most negatively affected by cap and trade, a question that explained briefly what framing is and asked them to name all the frames they could think of for the regulation of carbon dioxide emissions, and some generic usability questions.
Little is yet known about how frame reflection affects individual and group evaluation of competing positions, but several possibilities can be drawn from the literature (Cappella et al., 2002; Entman, 1993; Goffman, 1974; Hart, 2010; Price et al., 2005): frame reflection may encourage an understanding of and appreciation for competing arguments about a proposed policy; it may promote a better understanding of the policy's likely consequences for different groups of people; it may encourage people to imagine still other frames, and, accordingly, alternative positions on an issue. Our coding schema are based in part on this literature and in part on inductive analysis of sample responses. Inter-coder reliability was established on a sample of the coding categories using three coders' observations for 35 content coded categories, with Krippendorff's $\alpha=0.96$. While numerous aspects of participants' responses were coded, for clarity, we only describe here the aspects of the coding schema used in the subsequent analysis.

Participants in both conditions were asked to list as many frames as they could think of for the regulation of greenhouse gases. As an example, participants were told that health care is often framed in terms of the market (with the emphasis on health care users as consumers), in terms of rights (with health care depicted as an entitlement), and in terms of other frames.

We defined a frame as a shorthand way of talking about a policy in terms of the institution that would be most involved in its implementation or that is historically associated with policy solutions of that sort (Lau & Schlesinger, 2005). For example, a "market" frame would assess cap and trade in terms of its economic benefits. In coding for distinct frames, we used respondents’ own markers: for example, their enumeration of statements as "1, 2, 3," or their use of line breaks to demarcate statements. We suggest that participants who engaged in frame reflection would list a greater number of different frames.

We suggest further that respondents who had engaged in frame reflection would also identify frames drawn from a wider variety of source domains. By source domain we mean the institution associated with a particular justificatory rhetoric. For example, we coded the source domain of the frame, "people have a right to clean air" as Law since the justification is one of legal rights. Frames from an Economy domain used language associated with business, markets, industry, and economic interests. Those from an Environmentalism domain used language about the good of the environment, global protection, and so on. Other domains included Ethics (non-legal and non-environmental), Government, Medicine, and Science.
Analysis

Previous studies of framing have found that framing effects alone are often not as drastic as the interactions between framing other factors, such as previous opinions (Price et al., 2005). Therefore, rather than doing simple comparisons of the two conditions, our analysis examines which of the many factors involved are most predictive of the above described effects related to frame reflection.

To do so, we used all-subsets linear model selection with Mallow’s Cp to determine which other variables were most effective at predicting each outcome described above. This approach chooses the model that has the most explanatory power, in terms of fitting the data, and the least complexity, in terms of the number of predictors included.

Table 2 describes predictors used in model selection. Pairwise interactions were included for document count variables and for the condition with the Views principal components, since framing can interact with prior views and opinions (Price et al., 2005).

Table 2
Potential predictors and their interactions included in model selection.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CondVis / CondDocs</td>
<td>Experimental condition; visualization (1) or document reader (0).</td>
</tr>
<tr>
<td>Views PCs</td>
<td>(see Table 3)</td>
</tr>
<tr>
<td>Know PCs</td>
<td>(see Table 4)</td>
</tr>
<tr>
<td>LegTime</td>
<td>Time spent reading the legislative summary, in seconds</td>
</tr>
<tr>
<td>ToolTime</td>
<td>Time spent using the tool, in seconds</td>
</tr>
<tr>
<td>NumDocs</td>
<td>Number of documents viewed</td>
</tr>
<tr>
<td>NumCon</td>
<td>Number of Con documents viewed</td>
</tr>
<tr>
<td>NumMixed</td>
<td>Number of Mixed documents viewed</td>
</tr>
<tr>
<td>NumPro</td>
<td>Number of Pro documents viewed</td>
</tr>
<tr>
<td>NumDocs pairwise interaction with each of NumCon, NumMixed, NumPro</td>
<td></td>
</tr>
<tr>
<td>NumDocs pairwise interaction with each of NumCon, NumMixed, NumPro</td>
<td></td>
</tr>
</tbody>
</table>

To reduce the potential for multicollinearity, some predictors were collapsed using principle components analysis. Table 3 describes two principal components related to political ideology and views on global warming, capturing 74.6% of the variance in the four variables involved. Table 4 describes two principle components for prior knowledge about global warming and cap and trade.

Table 3
Principal component loadings for views related to global warming. The Views1 component is high for liberal individuals concerned about global warming, while the Views2 component is high for liberal individuals not concerned about global warming.

<table>
<thead>
<tr>
<th>Views1</th>
<th>Views2</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.312</td>
<td>0.699</td>
<td>Self-expressed political ideology; very conservative (1) to very liberal (5)</td>
</tr>
<tr>
<td>-0.457</td>
<td>0.698</td>
<td>Views on global warming, based on (Maibach et al., 2009); alarmed (1) to dismissive (6)</td>
</tr>
<tr>
<td>-0.476</td>
<td>0</td>
<td>Initial support for cap and trade, before using tool or reading legislative summary; strongly support (1) to strongly oppose (5)</td>
</tr>
<tr>
<td>-0.684</td>
<td>-0.157</td>
<td>Final support for cap and trade after using tool; strongly support (1) to strongly oppose (5)</td>
</tr>
</tbody>
</table>
Table 4
Principal component loadings for prior knowledge related to global warming. Know1 is high for individuals who feel uninformed about global warming but are in fact knowledgeable, while Know2 is high for those who feel uninformed and are indeed not knowledgeable.

<table>
<thead>
<tr>
<th>Know1</th>
<th>Know2</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Con</td>
<td>120</td>
<td>153,906</td>
</tr>
<tr>
<td>Mixed</td>
<td>69</td>
<td>107,124</td>
</tr>
<tr>
<td>Pro</td>
<td>38</td>
<td>49,205</td>
</tr>
</tbody>
</table>

This analysis excludes those participants who viewed zero documents, leaving 26 control (document reader) participants and 24 experimental (visualization) participants.

Results

Participants ranged age 18 to 21 (mean 18.8, median 19), with most describing themselves as moderate (36%) or somewhat liberal (32%) and as Democrats (40%), independents (16%) or Republicans (20%). Most were either “concerned” (46%) or “cautious” (32%) about global warming, resembling to some extent the distribution among general Americans (Maibach et al., 2009).

Time spent using the tool ranged from 10:04 to 25:06, with most taking slightly longer than the required 10 minutes (mean 13:35, median 11:43). Experimental participants spent slightly longer on average (14:10 > 13:02), but not significantly so (p=0.34). Experimental participants also saw fewer original documents than in the control condition (6.08 < 8.15, p=0.02), though the experimental condition also showed visualizations of patterns across all the documents.

Across both conditions, we saw some evidence for frame reflection. When asked to list framings for regulation of CO₂ emissions, participants suggested up to five different frames (mean 2.5, median 3), drawing on up to four different domains (mean 2.0, median 2). To reiterate, those participants who viewed zero documents are excluded here. As expected, we saw no main effect for use of the visualization on either number of frames or number of domains.

We next present the model selection results, i.e., which variables were most predictive of our indicators of frame reflection. Each model is described in terms of the coefficient and significance for each predictor (* p=0.05, ** p=0.01, *** p<0.001). Interactions are shown as variable names separated by a colon. These model selection results are both useful results themselves and help draw attention to relationships for subsequent exploration.

For the total number of frames participants suggested, the most predictive factors were better knowledge about global warming and viewing more documents. Since neither of these factors pertain to the experimental condition, we do not investigate this result further. However, the model for the number of domains on which participants’ frames drew (Table 5) shows two important results. First, the strongest and most significant predictor was participants’ prior knowledge about global warming; more knowledgeable participants drew on more source domains when suggesting new frames. Second, the experimental condition interacted with participants’ prior views in predicting the number of source domains.

Table 5
Model for the number of domains on which participants' new frames drew. $F_{6,43}=4.542, p=0.0012$.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Know2</td>
<td>-0.568 ***</td>
</tr>
<tr>
<td>NumMixedDocs</td>
<td>0.352 **</td>
</tr>
<tr>
<td>Views1:CondDocs</td>
<td>-0.183</td>
</tr>
<tr>
<td>Views1:CondVis</td>
<td>0.200</td>
</tr>
<tr>
<td>Views2:CondDocs</td>
<td>-0.115</td>
</tr>
<tr>
<td>Views2:CondVis</td>
<td>0.266</td>
</tr>
</tbody>
</table>
To examine this interaction further, we conducted ANCOVAs with Views1 and Condition, and with Views2 and Condition. For Views1, no significant effects were found. With Views2 and Condition, there were no main effects for either Views2 (F1,46=0.96, p=0.33) or Condition (F1,46=0.01, p=0.95), but there was a significant interaction effect (F1,46=7.11, p=0.011).

The Views1 and Views2 components differ primarily in their weighting for the Maibach et al. Maibach et al., 2009) global warming perceptions scale. Thus, we also examine the relationship between Maibach et al. segment and condition, finding that, again, neither segment (F1,46=0.05, p=0.82) nor Condition (F1,46=0.01, p=0.91) alone are significant, but their interaction is (F1,46=8.27, p=0.006) (see Figure 3). Specifically, in the control (document browser) condition, participants who were less concerned about global warming exhibited less evidence of frame reflection, whereas in the experimental (visualization) condition, those participants exhibited more evidence of frame reflection. Checking other factors in the model (Know2 and NumMixedDocs) indicates no other similar interactions.

Figure 3. Use of the visualization interacted with views on global warming (Maibach et al., 2009) in predicting the number of domains on which a participants' novel frames draw. In the control condition, those who thought global warming less of an issue drew on fewer domains, but the visualization condition reversed that trend.

In summary, two important findings emerge here. First, both prior knowledge and prior opinions about global warming were significantly related to indicators of frame reflection. Moreover, these prior views interacted with use of the visualization tool; in the control condition, participants less concerned about global warming drew on fewer source domains when suggesting new frames, but the visualization condition reversed this trend. Second, the relative proportions of Con, Mixed, and Pro documents also influenced participants’ thinking, a finding examined in more depth by subsequent follow-up analysis.
Follow-up Analysis

**Tool Usage.** To examine use of the visualization tool, we conducted further model selection analyses with data only from participants who used the visualization tool (N=24). From Table 2 above, we included those predictors that pertain to use of the tool (i.e., everything from ToolTime down). We also included the predictors and interaction in Table 6.

Table 6
*System usage variables used as predictors.*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NounsClicked</td>
<td>Number of unique nouns clicked on from the main list</td>
</tr>
<tr>
<td>VerbsClicked</td>
<td>Number of verbs expanded to show example fragments</td>
</tr>
<tr>
<td>ExsClicked</td>
<td>The number of example fragments clicked to show all similar fragments</td>
</tr>
<tr>
<td>VerbClicked:ExsClicked</td>
<td>– pairwise interaction</td>
</tr>
</tbody>
</table>

For the total number of frames suggested, the most important factor was the number of original documents to which participants were exposed (Table 7). However, this influence was mitigated by the number of example fragments on which the participant clicked. Essentially, only seeing example fragments was not as beneficial as seeing the source documents.

Table 7
*Model for which aspects of visualization usage predict number of frames. F\(_{2,21}\)=7.709, p=0.003.*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumDocs</td>
<td>0.592 ***</td>
</tr>
<tr>
<td>ExsClicked</td>
<td>-0.405 **</td>
</tr>
</tbody>
</table>

For the number of domains on which participants’ new frames drew, the single most important predictor was the number of documents viewed, though that was mitigated by the number of Pro and number of Con documents (Table 8). That is, viewing more total documents was beneficial, so long as there was not a preponderance of only Pro or Con documents. While not individually significant, the number of example fragments clicked contributed to fewer frame domains, suggesting a similar pattern as above.

Table 8
*Model for which aspects of tool usage predict number of framing domains. F\(_{5,18}\)=6.382, p=0.001.*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToolTime</td>
<td>-0.00141</td>
</tr>
<tr>
<td>NumDocs</td>
<td>0.994 ***</td>
</tr>
<tr>
<td>ExsClicked</td>
<td>-0.231</td>
</tr>
<tr>
<td>NumDocs:ProDocs</td>
<td>-0.0498 *</td>
</tr>
<tr>
<td>NumDocs:ConDocs</td>
<td>-0.0720 *</td>
</tr>
</tbody>
</table>

In summary, these results suggest that the visualization’s impact on frame reflection seems most noticeable when participants take advantage of the features that link them back to the original content to see the visualized patterns in context rather than looking only at examples of the pattern. Furthermore, we see additional evidence that the relative proportions of Con, Mixed, and Pro documents seem to be important predictors for our indicators of frame reflection. The next subsection examines this result further.

**Document Balance.** In the results described above, the relative proportions of documents expressing different views on cap and trade were found to predict frame reflection. First, we compare the proportions of documents viewed in each condition with the proportions in the data set of all documents by summing the counts of documents viewed in each condition (Table 9). Both conditions differed
significantly from the data set (Ctrl: $\chi^2=50.21$, df=2, $p<0.001$; Exp: $\chi^2=74.40$, df=2, $p<0.001$), and the two conditions also differed from each other, but not quite significantly ($\chi^2=5.69$, df=2, $p=0.058$).

Table 9
*Differences among the relative proportions of documents in the data set and documents viewed in each condition; counts summed across all participants.*

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Data Set</th>
<th>Ctrl (browser)</th>
<th>Exp (visualization)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Con</td>
<td>120 (52.9%)</td>
<td>65 (30.7%)</td>
<td>28 (20.1%)</td>
</tr>
<tr>
<td>Mixed</td>
<td>69 (30.4%)</td>
<td>45 (21.2%)</td>
<td>28 (20.1%)</td>
</tr>
<tr>
<td>Pro</td>
<td>38 (16.7%)</td>
<td>102 (48.1%)</td>
<td>83 (59.7%)</td>
</tr>
</tbody>
</table>

To examine these differences further, we performed stepwise model selection for three outcome variables, one each for the proportion of Con, Mixed, and Pro documents. Similar predictors were used as above, excluding those that deal with proportions of documents (NumConDocs and below in Table 5). We still include interactions between the Views principal components and the experimental condition. No model significantly predicted the proportion of Mixed documents. For both Pro and Con, there was a significant interaction between experimental condition and the Views1 component (model details omitted here in favor of ANCOVA). Focusing on the Maibach et al. (Maibach et al., 2009) segment, as above, reveals two important results. First, there is a significant main effect for condition, for both Con ($F_{1,46}=6.03$, $p=0.018$) and Pro ($F_{1,46}=4.83$, $p=0.033$) document proportions. Second, for Pro, there is a significant interaction ($F_{1,46}=6.08$, $p=0.017$) between condition and segment (Figure 4). Thus, in the control condition, participants exhibited something like an inverse selective exposure bias (Garrett, 2009; Munson & Resnick, 2010)—those who were less concerned about global warming read more Pro cap and trade documents, and vice versa. In the experimental condition, though, this trend was reversed, with those participants less concerned about global warming viewing more Con documents.

*Figure 4. Use of the visualization interacted with views on global warming (Maibach et al., 2009) in predicting the ratio of Pro documents a participant viewed.*
Discussion

The above results provide a detailed picture of how various aspects of the visualization influenced frame reflection. Here, we consider some broader ramifications.

First, the result in Figure 3—that the visualization may have helped support frame reflection among those who did not previously think global warming a serious issue but also reduced frame reflection among those who thought it was a serious issue—might be interpreted ambivalently, i.e., the tool helped some people's critical thinking and hurt others'. However, one might alternatively argue that the participants whose frame reflection was most impacted, those who did not consider global warming a serious issue, were the participants who could likely benefit most from increased frame reflection.

Second, we see evidence that aspects of the system design meant to bring users closer to the original data—clicking a verb to reveal an example fragment, clicking that example fragment to show similar fragments, and clicking a fragment to show it used in context—led to decreased frame reflection, unless the participant followed all the way through to the original document. This finding both aligns and contrasts with Baumer et al. (2010). Their study of metaViz found that focusing on example fragments pertaining to computationally identified metaphors decreased users’ critical thinking and creativity, even when the user did click through to the original document. This discrepancy could be explained in a variety of ways: metaViz is about metaphors, while the tool presented here is about framing; participants in the present study were given a less directed task when using the tool; the study of metaViz recruited Mechanical Turk workers while this study involved college students; embedding the content in the visualization, as was done here, might be more beneficial than linking to external content, as in metaViz. Future work using visualizations to promote critical thinking should attend to this tension between high-level patterns and low-level examples.

Third, the important role that the relative proportions of Con, Mixed, and Pro documents played, in addition to the differences in those proportions between the two conditions, raises the issue of selective exposure (Garrett, 2009; Munson & Resnick, 2010), i.e., seeking out information that resonates with one’s prior views and avoid conflicting information. The results here contribute to this area of study in two ways. First, they show that exposure to a variety of opinions helps facilitate frame reflection. Second, the results suggest that, while participants seemed to overcome their selective exposure bias in the control condition, the experimental visualization tool may have reinforced such biases (see Figure 4). Perhaps these participants had a general tendency toward diversity- or opposition-seeking that became too difficult to enact with the visualization. Since this experiment was not designed to examine selection exposure, though, it is difficult to determine exactly what might have caused the observed patterns. However, this result does draw attention to the need for consideration of how systems for reading political content, even those not expressly designed to broaden exposure (Munson & Resnick, 2010), may impact the variety of viewpoints to which users are exposed.

Future Work

Since the results presented here indicate that certain aspects of the design were associated with frame reflection, future studies should test different variants of such a visualization to determine more precisely the impact of each feature. For example, how might frame reflection by affected if the sidebar showing all example fragments opened as soon as a top-level noun was clicked, or if the full text of the original document was never shown?

Interpretation of the results should also be tempered by the study sample: students, mostly young, mostly liberal, mostly democratic. Furthermore, the modest sample size (N=50) may have been a factor in the model selection results. Future work should explore use of such tools among other populations, as well as considering a diversity of political issues.

While this study focuses on individuals, part of the motivation for this work is improving the quality of political deliberation. Future work should examine group settings, allowing deliberators to use the tool either before or while engaging in a deliberative discussion. It will be especially important to study non-laboratory settings, such as town hall style public forums or tools designed for readers of political blogs. Such work can help provide a better understanding of how people interpret, and make argumentative use of, the visualized linguistic patterns.
Finally, tools similar to that described here may also be useful for self-reflection, wherein the visualization is based on analysis of a user’s or group of users’ own discussion about a particular issue. Would using such a tool impact how the user thinks about her or his own perspective on an issue? Might a user subsequently change how s/he talks about, or thinks about, the issue? Although the work presented here focuses primarily on improving critical thinking about political issues, such a perspective could apply in a wide variety of areas.

**Conclusion**

This paper explores the design of an interactive tool to promote frame reflection (Schön & Rein, 1994), critical thinking about the ways in which a political issue is framed. In addition to describing the design of this novel system, we present results from a controlled lab study showing that use of the tool is linked to indicators of frame reflection. Specifically, when participants were asked to suggest novel frames, there was no difference in the number of frames suggested, but tool use was associated with drawing on more unique domains in considering alternative framings. Examining these results further, we find that users who took advantage of the tool’s ability to link higher-level linguistic patterns back to the original context in which they occurred exhibited greater levels of frame reflection. However, we also found evidence that study participants who used the visualization viewed slightly more documents that aligned with their own prior views than those who used a simple document browser. Thus while the tool seems effective at promoting frame reflection, it may also slightly reinforce selective exposure bias (Garrett, 2009; Munson & Resnick, 2010). This paper’s primary contribution, then, is not the details of the specific visualization evaluated here, but rather working toward building an understanding of how to design, as well as how to evaluate, tools for frame reflection. These results carry important implications in terms of implementing tools that not only facilitate political participation but encourage deep, thoughtful, and reflective engagement with the complex political issues facing our society.

**References**


An Examination of How the FCC Uses “Voluntary Commitments” from Merging Telecommunications Companies to Advance Policy Goals

Gwen Shaffer
California State University, Long Beach
gwen.shaffer@csulb.edu

Scott Jordan
University of California, Irvine
sjordan@uci.edu

Abstract

This study examines the FCC’s historical use of “voluntary commitments” when approving telecommunications company mergers. Because complex factors such as market conditions, corporate lobbying, political climate and technological change dictate regulations, it is grounded in a political economic framework. Using a focused synthesis, the authors examined key policy issues, such as the political climate and power structures in place during various telecommunications company transactions. The study contrasts the FCC’s ability to extract commitments from merging companies with previous unsuccessful attempts to achieve similar goals through the established rulemaking process, with particular focus on the 2011 Comcast/NBC-Universal merger. The newly formed company agreed to a slew of voluntary commitments that advanced policies—related to streaming video, digital inclusion and online journalism—strongly opposed by industry during previous FCC attempts to impose them industry-wide.

Keywords: information policy, qualitative data analysis, information services

The Federal Communications Commission (FCC) must justify significant policy decisions by explaining assumptions underlying the decision, the rationale for that particular policy tactic, and how the new policy dovetails with agency precedent (Yoo, 2007). Critics contend, however, that when federal regulators impose conditions on companies seeking to complete a merger, the process lacks transparency and is too far-reaching. In fact, the FCC may negotiate with companies under review to extract conditions that have minimal connection to actual concerns surrounding the transaction, and that circumvent established policymaking processes (Weiser, 2009; Noah, 1997; Tramont, 2001; Koutsky & Spiwak, 2010). Critics also assert that by engaging in closed door negotiations and “arm-twisting” (Noah, 1997), the FCC is able to evade judicial scrutiny—partially because companies in regulated industries fear repercussions if they challenge agency demands. For these reasons, members of Congress have sponsored legislation aimed at severely limiting, and even stripping, the FCC of its power to review acquisitions (Borland, 1999; Telecommunications Merger Review Act of 2000).

Despite their controversial nature, merger negotiations often do result in a set of conditions that benefit the public. It may be argued that these outcomes would be impossible without the agency’s “sector-specific” (Baker, 2010) perspective. This outlook is shaped by the FCC’s industry expertise and ongoing relationship with telecommunications companies that offer broadband access, voice-over-IP and digital content—in contrast to the limited interaction between the Department of Justice (DOJ) and merging firms. While the DOJ tends to pore over documents and crunch numbers, the FCC review process entails meeting with the parties and hearing their concerns directly. In a converged environment, in particular, merger conditions offer an interim solution for the FCC to work outside an antiquated regulatory scheme (Blumensaadt, 2000) and to spark broader regulatory reform. Negotiated terms are a

Copyright is held by the authors.
quick route compared to the FCC’s established rulemaking process and may serve as a useful pilot. Former FCC Chairman William Kennard (1999) lauded the practice as a means of meeting the public interest standard, while freeing the FCC from the need to predict how rules might impact every company in the sector.

Whether the commission’s authority to condition mergers benefits or hinders regulation remains debatable. With this sentiment in mind, this study examines the FCC’s historical use of “voluntary commitments” when approving telecommunications company mergers. Because complex factors such as market conditions, corporate lobbying, political climate and technological change dictate regulations, this study is grounded in a political economic framework. Using a focused synthesis, the authors examined key policy issues, such as the political climate and power structures in place during various telecommunications company transactions. The study contrasts the FCC’s ability to extract commitments from merging companies with previous unsuccessful attempts to achieve similar goals through the established rulemaking process. While multiple acquisitions are analyzed, the Comcast/NBC-Universal merger serves as a key case study. The newly formed company agreed to a slew of voluntary commitments that advanced policies—related to streaming video, digital inclusion and online journalism—strongly opposed by industry during previous FCC attempts to impose them industry-wide.

Current Regulatory Review Process

When reviewing mergers and acquisitions of telecommunications companies, the DOJ’s primary objective is ensuring communications markets perform competitively (Baker, 2010). By contrast, the Communication’s Act charges the FCC with determining whether proposed transactions meet a “public interest” standard. This broad mandate may encompass protecting service quality for consumers, preserving American jobs, safeguarding localism, or providing opportunities for audiences to hear diverse points of view (FCC, 2004). When a disagreement arises during an acquisition and license transfer, the FCC may refer the matter to an administrative law judge. Far more commonly, however, the key stakeholders “voluntarily” agree to narrowly tailored conditions geared toward mitigating harm and ensuring the transaction under review does not violate FCC rules. Together, three general principles drive the FCC’s information policy: civil liberties, economic efficiency, and social fairness and equity (Bushkin & Yurow, 1980). The following section briefly reviews key literature examining the political economy of telecommunications and the FCC’s public interest mandate.

Political Economy and Public Interest Literature

It is impossible to view the telecommunications industry in isolation from regulators and lawmakers, or to view regulators and lawmakers as separate from the market. Each stakeholder is linked to another. Because these relationships have a profound impact on consumers, the Communications Act mandates regulators to consider another key concept—the public interest. Consequentially, political economy provides an explanatory framework for analyzing both policy and practices influencing telecommunications mergers.

Ideally, telecommunication regulations should promote economic, political and social processes—which then enable a democratized society (Lenert, 2006). Although the FCC accepts public comments before signing off on a merger, these comments may not present the type of information the commission relies on to justify its opinions (Blevins & Brown, 2010). Powerful companies and those with the highest financial stakes routinely usurp the debate over telecommunications policy in the United States (Chen, 1997). Certainly, gaining traction in the hegemonic U.S. telecommunications industry, where a handful of companies control wireline and wireless Internet access, is a challenge. Verizon, AT&T, Comcast, and Time Warner spent a combined $53.3 million to lobby Washington policymakers in 2010 (Center for Responsive Politics, 2011). This large sum of money helps explain why legislators sometimes craft a regulatory framework that privileges the interests of corporations over the interests of broadband consumers.

When a small number of elite and politically connected ISPs establish the terms and conditions for access to the Internet and digital content, by default these firms also determine who can operate in the public sphere. The Supreme Court acknowledged the concepts of a free marketplace of ideas and media diversity as far back as 1945. In the landmark case Associated Press v. United States (1945), the justices
ruled that “the widest possible dissemination of information from diverse and antagonistic sources is essential to the welfare of the public” (p. 17). Similarly, the Communications Act of 1934 recognizes the importance of fostering competition, on the grounds that it leads to innovation, lower prices and better quality service. Inversely, telecommunications industry consolidation has the potential to undermine quality of service, limit consumer choice, and lead to price spikes. Previous research has found that if the FCC over-emphasizes the free market when crafting policy, it could dampen deliberative discourse and citizen engagement (Blevins & Brown, 2010). However, “free enterprise” economists argue that government intervention itself is harming competitive markets (Koutsky & Spiwak, 2010; White, 2008). Certainly, when regulatory requirements are applied to select companies, a patchwork of inconsistent policies begins to comprise the political economy of telecommunications. The SBC-AT&T and Verizon-MCI mergers, which closed in 2005 and 2006 respectively, serve as examples. These companies agreed to comply with net neutrality principles for two years. Yet regional DSL providers and the national cable industry, which boasted more residential broadband subscribers than AT&T and Verizon combined, were not bound by these same restrictions.

Technology is enabling users to exercise more control over media platforms, from audio filesharing to video-on-demand. As a result, the political economy of telecommunications is becoming more user-driven, and telephone and cable companies are reconsidering their business models. It remains to be seen how these changes will impact conditions placed on mergers in the future, or on the FCC’s ability to enforce them. However, regulators and policy scholars continue to learn from voluntary commitments agreed to in the past. Three research questions emerged from this review of literature:

- **RQ1:** Do merger conditions appear to be tailored to the material facts of the transaction under review by the FCC?
- **RQ2:** Are merger conditions driven by regulators’ broad policy agenda? Specifically, do these “voluntary commitments” move the FCC closer to achieving policies overwhelmed by opposition during the traditional rulemaking process?
- **RQ3:** When crafted appropriately, can merger conditions transform an anti-competitive merger into a transaction that achieves a net gain for the public?

The following section explains the methodology used to explore and, ultimately, answer these questions.

### Methodology

This qualitative study uses empirical evidence to shed light on processes and outcomes surrounding telecommunications company mergers. The authors conducted a focused synthesis (Majchrzak, 1984) involving the review of regulatory policies, scholarly research findings, white papers, and other written materials. Given the vast amount of available information, the researchers selectively chose relevant documents by applying both external and internal criticism to each. External criticism questions whether the document is authentic, and internal criticism questions whether the data are accurate and relevant (Isaac, Stephen, & Michael, 1981). Similarly, we considered whether these original sources were reliable and valid (Cooper & Emory, 1995), and whether biases were evident.

The Telecommunications Act of 1996 set the stage for the telecom consolidation trend that continues today. Therefore, the first step we took involved examining the dozens of telecommunications company mergers that have occurred since its passage. We then identified the research problem—broadly, what drives specific merger conditions and whether they compensate for a less competitive marketplace. Next, we focused on the policy environment, i.e. the political climate and power structures in place, during various transactions. This included an examination of overt stakeholder actions, such as lobbying efforts by corporations and grassroots advocacy. We then identified pending FCC rulemakings, regulatory proposals and legislation that addressed issues overlapping with merger conditions. Based on this information, we considered the goals of various merger conditions, how they evolved, and their effectiveness. We then analyzed whether they ultimately served the public interest. Finally, we synthesized these findings to answer our research questions.

The following section examines a handful of telecommunications company transactions, and draws connections between the political agenda, the marketplace and merger conditions.
Historic Use of Merger Conditions

Across all industries, firms are motivated to merge with or acquire other companies to gain a competitive advantage. By joining forces, telecommunications companies aim to strategically combine resources, to maximize efficiency, to obtain new customers, and to extend services into new markets. During the 1990s, America Online (AOL) grew into a preeminent Internet company. In addition to creating a software suite that allowed its 30 million members (Holahan, 2006) to access the world’s largest walled garden of content and community forums, AOL operated two ISPs and various web services. In January 2000, AOL proposed buying Time Warner for $165 billion. As the largest provider of content in the world (MacKie-Mason, 2000), Time Warner operated multiple cable television systems and popular networks including HBO, Cinemax and CNN; sports franchises; magazines; film production and distribution companies; and record labels. Time Warner cable also owned the nation’s second largest Internet service, branded Road Runner. The proposed transaction exemplified the converging environment between “new” and “old” media. Time Warner recognized that the Internet was crucial to the future success of its music, publishing and video businesses. AOL, in turn, hoped to gain access to Time Warner’s vast content collection, as well as to its Road Runner broadband service.

Despite the companies’ claims that their businesses did not overlap enough to warrant intense scrutiny, the FCC spent nearly 10 months conducting a public interest analysis. Ultimately, both the FCC and the FTC approved the transaction with about a dozen conditions. Time Warner was required to provide at least three unaffiliated ISPs with access to its cable systems. The order also prohibited AOL Time Warner from discriminating against content delivered to subscribers of competing ISPs but traveling over its cable system. Finally, AOL Time Warner agreed to make future instant messaging (IM) services interoperable. In aggregate, these conditions advanced the FCC’s long-held goal for cable broadband providers to offer “open access” to their infrastructure (FCC, 2000). At the time, the telephone system operated with full common carrier obligations. This meant that competitive ISPs could sell services within a community for minor investment, with the telephone company providing the last-mile connection to its customers. In the absence of common carriage obligations for cable operators, however, a handful of companies controlled broadband access. In response to RQ2, the FCC capitalized on the AOL-Time Warner merger as an opportunity to mandate that a cable company allow unaffiliated ISPs access to its broadband network. In response to RQ3, our analysis finds that the FCC (2002; 2003) ultimately negated potential benefits of the AOL-Time Warner merger by lifting this requirement, as well as the mandate to make IM services interoperable.

The AOL-Time Warner transaction was reminiscent of AT&T’s $43.5 billion acquisition of Tele-Communications, Inc. (TCI), approved by the FCC just one year earlier. That 1999 deal allowed AT&T—the nation’s largest telephone carrier—to combine its consumer long distance, wireless, and Internet services with TCI’s cable, telecommunications, and high-speed Internet business. At the time, TCI was the nation’s second-largest cable television operator. However, the FCC did not mandate that the newly formed entity provide competing broadband companies with access to TCI’s cable lines. We argue that this contradiction is solely the result of political priorities, as opposed to the material facts surrounding the merger. Specifically, the FCC undertook this merger review before the open access debate gained traction. Instead, cable television and telephone competition; a V-chip order; broadcast ownership rules; universal service; and the transition from analogue to digital television dominated the commission’s agenda (Kennard, 1998).

AT&T acquired additional regional telecommunications companies over the next few years. For instance, a 2006 horizontal merger allowed AT&T to purchase BellSouth for $86 billion. In order to win approval from federal regulators, the new company signed off on 11 pages of merger commitments. AT&T agreed to adhere to net neutrality principles for two years and to divest BellSouth’s holdings in the 2.5 GHz spectrum (FCC, 2006). These commitments directly address federal regulators’ concerns about competitiveness—always the primary issue during a merger. Several stipulations, however, were wholly unrelated to the terms of the deal (RQ1) but address broader federal policy goals (RQ2). For instance, AT&T agreed to make disaster recovery capabilities available in BellSouth’s territory and to donate $1 million toward supporting public safety initiatives. Several months earlier, an independent panel established by the FCC in the wake of Hurricane Katrina recommended ways in which the telecommunications industry could more effectively respond to disasters (Wiley, Rein & Fielding, 2006). The merger agreement also required AT&T/BellSouth to report to the FCC on its efforts to serve.
customers with disabilities. The commission had struggled to address a variety of disability-related matters—including access to telecommunications equipment, services and information—since passage of the Disabilities Rights Act in 1990. Additionally, in order to win approval for its merger with BellSouth, AT&T agreed to return outsourced jobs to the United States. This condition is in line with the federal government’s general commitment to job creation and economic growth. Finally, the FCC required that AT&T/BellSouth slash rates charged to competitors wishing to lease high-speed data lines. This condition allowed the FCC to claim a victory in its effort to reform special access fees, which it could not implement industry-wide due to opposition.

AT&T’s 2006 acquisition of BellSouth has much in common with the recent horizontal merger between CenturyLink and Qwest. The FCC spent a year reviewing the $12.2 billion deal before approving it in April 2011, and creating the nation’s third largest telecommunication provider (CenturyLink, 2011). The commission imposed conditions on Qwest and CenturyLink that reflect a series of goals laid out in the FCC’s National Broadband Plan (2010). The plan stresses the need for affordable and reliable residential broadband access, with an emphasis on rural, low-income and minority communities. In order to win merger approval, CenturyLink and Qwest agreed to expand broadband access to low-income customers in their territory. They also promised to offer qualifying households broadband access for $10/month for one year, and $15/month for the subsequent year. Additionally, CenturyLink agreed to increase Qwest’s network capacity so that faster broadband download speeds would be available to 4 million additional homes and businesses. Each of these merger conditions explicitly addresses proposals laid out in the National Broadband Plan, demonstrating that a broad regulatory agenda frequently drives merger conditions.

The following section, which focuses exclusively on Comcast’s recent acquisition of NBC-Universal, further explores the answer to our research questions.

Comcast/NBC-Universal Merger Conditions

Motivations for the Merger

In January 2011—following six congressional hearings, multiple public forums, and one of the lengthiest comment periods in commission history—the FCC sanctioned Comcast’s purchase of NBC/Universal. This $30 billion transaction illustrates the relationship between merger conditions and the FCC’s existing policy goals, including many the commission had failed to achieve through administrative rulemakings. In light of the role campaign finance contributions play in determining the political economy of telecommunications, this section also examines lobbying efforts undertaken by both supporters and opponents of the Comcast-NBC/Universal deal.

More than 24 million Americans subscribe to Comcast’s cable TV service, making it the largest multichannel video programming distribution service in the United States. With 15 million broadband subscribers, Comcast is also the largest ISP in the country. However, like all cable companies, Comcast’s traditional business model is threatened by online video and growing competition from satellite and phone companies that offer subscription video services (i.e. Verizon FiOS). Comcast recognized the necessity to transition from a subscription-based cable company to a media content provider. It stands to profit from NBC-Universal’s holdings (Associated Press, 2009), which include the NBC and Telemundo broadcast networks; 26 local TV stations; cable channels including CNBC, Bravo and Oxygen; the Universal Pictures movie studio and theme parks; and a 30 percent stake in online video distributor Hulu.com (Associated Press, 2011).

The Political Economy of the Merger

After separately reviewing the Comcast/NBC-Universal merger for nearly a year, both the DOJ and the FCC approved it in January 2011. The final consent decree contains multiple voluntary commitments that, we argue, are meant to advance a policy agenda pushed by public interest groups and FCC leadership (RQ1 and RQ2).

While the FCC makes evidence-based decisions after reviewing public comments and considering the public interest, its decisions are also “politically informed” (Cowhey, Aronson, & Richards, 2009, p. 108). Policy outcomes are, often, the direct result of the clout exhibited by each stakeholder
attempting to influence the political process through contributions to political campaigns, mobilizing voters and swaying public opinion (Galperin, 2004). While working to gain approval of the merger, Comcast had more than 30 lobbying firms on its payroll and spent about $100 million. Nearly 100 House of Representatives members signed onto a letter urging the commission to approve the Comcast/NBC-Universal merger in order to promote “jobs and investment” (U.S. Congress, 2011). According to the Center for Responsive Politics (2011), 84 of the letter’s 97 signatories accepted political contributions from Comcast, ranging from about $1,000 to $25,100. All told, the cable company made political contributions to 385 members of Congress—75 percent of the legislative body—serving at the time of the merger negotiations (Free Press, 2011; Nichols & McChesney, 2011). The large sums of money spent on lobbying help explain why legislators might craft a regulatory framework or sanction a transaction that privileges the interests of corporations over those of consumers (Birdsall, 1996).

At the same time, multiple aspects of the Comcast/NBC-Universal merger suggest the political economy is shifting. Public interest groups opposed to the deal floated petitions, released fact sheets, posted to blogs, testified on Capitol Hill, and placed full-page ads in Beltway publications. They cautioned the merger would enable Comcast to slow down or block streaming content from competitors such as Netflix and iTunes; allow Comcast to raise programming and service rates; and shrink an already scarce supply of independent and diverse voices on television. Yet when federal regulators approved the Comcast/NBC-Universal merger on January 18, 2011, even some of the deal’s staunchest critics applauded conditions placed on the newly formed company (Crawford, 2011; Feld, 2011; Consumer Federation of America, 2011). In fact, Comcast/NBC-Universal ultimately agreed to a slew of voluntary commitments that advance a policy agenda advocated by public interest groups—from expanding the marketplace for online films and television shows and supporting digital journalism, to requiring the company to offer ethnic programming and lower-cost broadband.

A Vehicle for Achieving a Broad Policy Agenda

In its analysis of the Comcast/NBC-Universal transaction, the FCC expressed concerns that the joint venture could boost the prices that competing video distributors pay for the right to distribute NBC-Universal programming, or make it possible for the new company to sell only unpopular shows and movies to competitors (Stoltz, 2011). These scenarios would result in customers switching to Comcast in order to gain access to the shows they value, and then getting locked into costly contracts. Therefore, one of the merger conditions allows an online video distributor to pursue a commercial arbitration procedure, should it become entangled in a licensing dispute with Comcast (DOJ, 2011).

We speculate that clashes between cable operators and programmers inspired this particular merger condition. In October 2010, News Corp. doubled the annual rate it charged to carry the Fox network on several New York and Philadelphia channels to $150 million. Cablevision refused to pay up and, as a result, 3 million Cablevision customers temporarily lost access to Fox programming. News Corp. also blocked Cablevision Internet customers from watching Fox content on both the network’s website and on Hulu. A similar feud over retransmit fees caused DirecTV customers in a handful of markets to lose Fox programming in January 2011. Around that same time, Time Warner Cable and Sinclair Broadcasting resolved a lingering fight over retransmission fees affecting 28 stations. Even in blackout situations, the FCC can do little beyond reminding the parties of their obligation to negotiate retransmission agreements “in good faith.” The arbitration clause in the Comcast/NBC-Universal merger consent decree opens the door for the FCC to implement a more sweeping policy, which commissioners were crafting at the time of the merger approval (FCC, 2011).

The FCC imposed another merger condition requiring Comcast to offer standalone broadband service for less than $50 per month for three years. This condition may be an outgrowth of “lessons learned” during previous mergers. In approving both the SBC-AT&T and Verizon-MCI mergers in 2005, the FCC mandated these carriers sell stand-alone DSL service for two years. But, allegedly, the companies priced naked DSL service so high that few customers purchased it without also buying phone service (Consumers Union, 2006). By capping the fee Comcast may charge, the FCC is attempting to ensure customers are not forced to subscribe to unwanted video and phone services. Because the commission has pushed for such assurances previously, this merger condition speaks to RQ2. Because this condition clearly serves the public interest, it also helps answer RQ3.
The “Internet Essentials” program incorporated into the merger agreement ensures that every household in Comcast’s footprint with children eligible for the federal free lunch program qualifies to receive broadband service for $10 per month. Internet Essentials also provides PCs for $150 and access to digital literacy training. This condition does not address antitrust concerns that typically dominate merger reviews, but it does directly advance the FCC’s digital inclusion goals—answering our first two research questions. Specifically, the FCC’s National Broadband Plan (2010) recommends “free or very low-cost wireless broadband” as a means to spur broadband adoption. The plan also cites the need to teach digital literacy skills and make non-adopters more comfortable with computers. By tacking on these conditions, the FCC also created a public good (RQ3).

As political economists McChesney and Schiller (2003) assert, communication systems are crucial for democracy because they provide a space for citizens to clarify and mobilize around social priorities. In its agreement with the FCC, Comcast/NBC-Universal pledged to establish partnerships between non-profit news organizations and at least five NBC-owned television affiliates. Similar online publications with a focus on “hyperlocal” news have popped up all over the country, typically in response to downsizing at mainstream newspapers. For instance, the Seattle PostGlobe features blog posts and articles written by citizen journalists. The joint ventures mandated by the Comcast/NBC-Universal consent decree are meant to be modeled after the Voice of San Diego—an online news organization that partners with the local NBC-owned TV station.

While a dearth of local news coverage is an emerging problem (Waldman, 2011; Hindman, 2011; Purcell, Rainie, Mitchell, Rosenstiel & Olmstead, 2010), it is only tangentially relevant to the marriage between Comcast and NBC-Universal. This commitment does, however, reflect a key priority for the FCC and some federal lawmakers. In 2009, the Senate held hearings focused on “the future of journalism” and introduced legislation that would allow newspapers to become tax-exempt non-profits. In 2010, the FTC hosted two workshops examining journalism in the Internet age. The Federal Trade Commission (2010) also proposed policies intended to support the “reinvention of journalism.” In 2010, the FCC’s charged its “Future of Media” task force with examining how news organizations serve local information needs of citizens. Clearly, the commitment made by Comcast/NBC-Universal to support non-profit news is meant to advance a broader policy goal articulated by federal policymakers (RQ2) and to sustain outlets for deliberative discourse (RQ3).

At least one other aspect of the consent decree illustrates that some merger conditions are driven by broad policy goals, rather than based on concerns related to a particular transaction. As part of its agreement with the FCC, Comcast/NBC-Universal promised that 10 NBC owned and operated stations would produce an additional 1,000 hours of original, local news programming, and that Telemundo stations would air a new Spanish-language multicast channel. In fact, the FCC began increasing its focus on Spanish language programming years earlier. In 2007 and 2008, the FCC adopted new rules for children’s television programming. In 2008, Congress passed the Child Safe Viewing Act, requiring the FCC to support the development of advanced content blocking technologies.
Conclusion

Through the use of focused synthesis, combined with a political economic analysis, we were able to definitively answer our first two research questions. Multiple examples demonstrate that voluntary commitments frequently advance regulators’ broad policy agenda, and that they are not tailored to the transaction under review. Our analysis also identified numerous instances in which the FCC attempted to achieve a policy goal that faced overwhelming opposition when pursued through the traditional rulemaking process. Finally, RQ3 asks if thoughtfully crafted merger conditions can transform an anti-competitive transaction into a net gain for the public. Definitively answering this question is more difficult. We find that the potential for public good exists, but the ability to transform a troubling acquisition depends on the FCC’s willingness to enforce conditions, as well as on market forces. For instance, Comcast/NBC-Universal agreed to significant conditions meant to facilitate competition with online video distributors. The FCC is banking on that fact that services like Netflix, iTunes, Hulu and Google TV will be on a level playing field with cable companies by the time the condition expires in 2018. If the market for streaming video shifts before then, consumers win and the FCC’s public interest goals are met. If the status quo remains unchanged, however, one of the merger’s most negative impacts will simply be delayed for seven years.

References


Federal Communications Commission (2000). Inquiry concerning high-speed access to the Internet over cable and other facilities, GN Docket No. 00-185.


Federal Communications Commission (2003). Applications for Consent to the Transfer of
Control of Licenses and Section 214 Authorizations by Time Warner Inc. and America Online, Inc., Transferors, to AOL Time Warner Inc., Transferee; Petition of AOL Time Warner Inc. for Relief From the Condition Restricting Streaming Video AllHLS, 18 F.C.C. Rcd. 16835 ¶ 6 (2003).


of Communication, 48(4), 3-23.


S. 2825 (December 3, 2009). Cell phone early termination fee, transparency, and fairness act. 111th Congress.


U.S. Congress (2011, January 5). Correspondence with FCC Chairman Julius Genachowski. MB Docket No. 10-56.


Markets Center, working paper 08-04.
Buying numbers: An Empirical Analysis of the IPv4 Number Market

Milton L. Mueller
Syracuse University
Mueller@syr.edu

Brenden Kuerbis
Syracuse University and University of Toronto
bkuerbis@gmail.com

Abstract

The emergence of a trading market for previously allocated Internet number blocks is an important change in Internet governance. Almost all of the Internet’s 32-bit address space has been given out, and we have not migrated to a new internet protocol, IPv6, with a larger address space. IP addresses are therefore being commoditized, as organizations with surplus numbers sell address blocks to organizations that want more. Though controversial, we know little about this phenomenon. This paper quantifies the number of address blocks that have been traded as of July 2012 and analyzes the scant information that exists about the pricing of these resources, discovering the emergence of a billion dollar market. The paper then shows how this factual information relates to key policy debates, in particularly the role of needs assessment and property rights in IPv4 number blocks.

Keywords: Internet governance, IP address markets, IPv4 scarcity, IPv6 migration, property rights

Introduction

One of the most important but least-studied aspects of Internet policy is the emergence of a trading market for previously allocated Internet number blocks. Without unique Internet protocol numbers for the networks and devices attached, the Internet simply doesn’t work. The original Internet Protocol standard, known as IPv4, specified a 32-bit numbering space, which provided slightly less than 4 billion unique numbers that could be used as addresses (Postel, 1981). A large part of that number space has already been handed out to organizations. The available supply is dwindling. The Asia-Pacific region is already reduced to rationing its last /8 in tiny 1024-address chunks, one to an organization. The European region is only a few months from that status.

Recent market developments and policy changes by Internet number registries now allow organizations with more numbers than they want to sell them to another organization. In other words, a market for IPv4 numbers is now possible. This is a major change in the political economy of Internet governance. IP numbers were considered common pool resources and registry contracts allocating number blocks to users made them forswear any property rights in their blocks. It is likely that the commercial forces unleashed by number markets will have far-reaching consequences for Internet businesses, users and governance institutions. Certainly domain name registration was transformed by the emergence of a commercial market, leading to the growth of a domain name industry and major changes in policies and institutions, such as the formation of ICANN, the separation of registries and registrars and the new gTLD program.

Number markets are controversial. Many Internet traditionalists resist the commoditization of what they view as a “community resource.” Others see it as diverting progress toward IPv6. Still others predict that “a functioning market won’t form at all, or will break down very quickly after it forms” (van Beijnum, 2011). This paper makes a much-needed empirical contribution to the literature on the economics and institutions of IP addressing. While several papers already discuss IP number markets in theoretical or


Copyright is held by the authors.
policy terms, no one has actually compiled and analyzed the transactions themselves. This paper draws on the records of the RIRs to compile as much information about traded IP number blocks as possible, and then conducts some very basic analysis of stocks, flows and proportions to assess the nature of this emerging market and explore some of its implications for Internet governance.

**Background: The Controversy and the Policy Changes**

This section explains the significance of what are known as “legacy” or “historical” allocations in the emerging market, and then describes the policies the RIRs passed authorizing market transfers and the date they were passed.

**Legacy Number Allocations**

Prior to 1991, there were no regional registries with formal policies for allocating number resources. There was only a central registry known as the Internet Assigned Numbers Authority (IANA), run by USC’s Information Sciences Institute (Cerf, 1990). Upon receipt of number resources from IANA, organizations did not have to sign contracts governing their use. Conservation was not a major consideration in this early period of Internet development. More than 40% of the IP address space was given away as universities, the U.S. military and major corporations involved in data networking projects received large allocations simply by asking for them. Many of them retain those allocations to this day. Organizations that received IP number blocks prior to 1991-1999 are considered “legacy” or “historical” holders, and their number holdings are usually not subject to RIR contracts created later.

**IPv4 Scarcity and the IPv6 Transition**

As Internet usage increased, and after the RIRs assumed responsibility for managing IP numbers, increasingly stringent conservation policies were applied (Hubbard, Kosters, Conrad, Karrenberg, & Postel, 1996). To receive an allocation one had to “demonstrate need,” which meant providing the RIR with confidential business and technical information about how the requested numbers would be used. To gain new allocations one had to demonstrate a specific utilization level of one’s existing allocation. In theory, numbers that were not needed were supposed to be given back to the RIR. In practice, few number blocks were returned. Those holding number blocks had no incentive to return them, and the RIRs lacked the institutional capacity or authority required to take them back (Mueller, 2010). The combination of weak-to-nonexistent reclamation and the delegation of 40% of the IPv4 numbers to legacy holders meant that despite the apparent scarcity of IPv4 numbers, there were many unused or underutilized number blocks (Mueller, 2008; Perset, 2007).

Some say that the whole problem of IPv4 scarcity will go away because there is a new Internet protocol with an expanded address space. This is a half-truth, at best. It is true that by 1998 Internet engineers had developed a new Internet protocol, IPv6, with a gargantuan, 128-bit address space. But in one of the most fateful and questionable design decisions in the IETF’s history, the IPv6 protocol was not made backwards compatible with IPv4. The incompatibility means that to implement IPv6, one must either abandon communication with everyone else running IPv4, or run both protocols at the same time. And running both protocols in parallel (known as ‘dual stacking’) does not reduce the demand for IPv4 numbers. So the migration from IPv4 to IPv6 does not involve an incremental reduction in the demand for IPv4 numbers as networks adopt IPv6. Rather, it involves parallel growth of the IPv4 and IPv6 Internets until such time as a huge tipping point is reached. When almost everyone is running IPv6 then, and only then, can the networks running IPv6 shut off IPv4.

---

1 The date varies because RIRs were established in different regions at different times. ARIN wasn’t created until 1997, whereas in Europe RIPE-NCC was established in 1991 and had a contractual governance scheme in place by 1994.
Transfer Policies

Beginning in 2005 the Internet technical community began to notice that the supply of unallocated IP numbers was nearing exhaustion (Hain, 2005). Migration to the new IPv6 protocol still seemed years away (Colitti, Gunderson, Kline, & Refice, 2010; Dell, Kwong, & Liu, 2008; Elmore, Camp, & Stephens, 2008). That is when the idea of number markets began to be seriously considered. So controversial was the idea of trading address blocks that the terminology used by the RIRs studiously avoided explicitly stating that numbers would be bought and sold. They were called “transfers to specified recipients” or “transfers of allocations.” Advocates claimed that market transfers would provide incentives for blocks holders with underutilized allocations to make them available to companies that needed them. The rising price of increasingly scarce IPv4 blocks would encourage a gradual and economically rational migration to IPv6. Others feared that markets would be so successful at extending the life of IPv4 that it would become a roadblock on the path to IPv6. This debate is covered in other literature and will not be recounted here (Edelman, 2008; Hofmann, 2010; Lehr, Vest, & Lear, 2008; Mueller, 2008). Suffice it to say that in most regions market transfer policies were passed after long, rending debates.

RIPE-NCC, the RIR for the European region, was first to approve a policy authorizing commercial transfers of IP address blocks in 2008. The RIPE NCC also runs a “Listing Service” that allows holders of IP address blocks to advertise numbers available for “exchange.”2 ARIN, the RIR for the North American region, implemented a market transfer policy in June 2009, and also runs a listing service.3 APNIC, which serves the Asia-Pacific region, was one of the first to propose a transfer policy, but debate and controversy prevented its adoption until February 2010. The other two RIRs have not yet authorized market transfers of IPv4 address blocks. Table 1 below summarizes the situation and contains links to the relevant policy documents governing market transfers.

Table 1
RIR Market Transfer Policies

<table>
<thead>
<tr>
<th>Region</th>
<th>Transfer policy</th>
<th>Date passed</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIPE NCC (Europe)</td>
<td><a href="http://www.ripe.net/ripe/docs/ripe-553#----transfers-of-allocations">http://www.ripe.net/ripe/docs/ripe-553#----transfers-of-allocations</a></td>
<td>December 2008</td>
</tr>
<tr>
<td>APNIC (Asia-Pacific)</td>
<td><a href="http://www.apnic.net/policy/transfer-policy">http://www.apnic.net/policy/transfer-policy</a></td>
<td>February 2010</td>
</tr>
<tr>
<td>LACNIC (Latin &amp; Carib)</td>
<td>No transfer policy in place</td>
<td></td>
</tr>
<tr>
<td>AFRINIC (Africa)</td>
<td>No transfer policy in place</td>
<td></td>
</tr>
</tbody>
</table>

Market and Non-market Transfers

We define market transfers as those transfers authorized by the new transfer policies listed above. More generically, they are an agreement between separate organizations to transfer the registration of a specified IP number block from the holder to the recipient in exchange for money. Market sales are not the only way to transfer IP address allocations from one organization to another. Transfers also occur through mergers and acquisitions. There are also specified methods by which organizations with multiple subsidiaries or business units with different corporate accounts with the RIR transfer number resources among themselves (internal transfers). We refer to internal transfers and transfers arising from mergers and acquisitions as “non-market” transfers. In order to study the emerging transfer market, it is necessary to differentiate records of non-market transfers from records of market transfers. Some RIR data regarding transfers, notably APNIC, combine logs of non-market transfers and market transfers and do not differentiate between them (see Data and Method section below).

2 RIPE NCC Listing service: https://www.ripe.net/lir-services/resource-management/listing. As of August 1, 2012 it showed 4 /21s (8,192 numbers) available.
3 ARIN specified transfer listing service: https://www.arin.net/resources/transfer_listing/index.html
Data and Method

There is no single source of data about market transfers of IPv4 number blocks. This section describes how the authors cobbled together the data used in this paper, drawing on bankruptcy court records, RIR statistics, Whois records, and RIR lists of transferred blocks.

The IP address Whois is a database that shows the general public which organization is holding which IP address block(s). Each RIR operates its own Whois database. It provides a tool for network operators and others to discover which number blocks are allocated or assigned to which organization. Using web-based interfaces, one can type the name of an organization and see a list of the IP number blocks registered in its name, or one can type in an IP address and see which organization has registered the block containing it. It may also be possible for an organization to obtain “bulk access” to the Whois records of an RIR, which means that the entire database is downloaded and can be processed and analyzed automatically. At least one RIR also maintains a “Whowas” database, showing which organization held a number block at specified dates in the past.

The 3 RIRs with transfer policies take entirely different approaches to recording number transfers. RIPE publishes no information whatsoever about events that occur under its market transfer policy. ARIN publishes a list of IP Address Blocks Transferred per NRPM 8.3. The list, updated weekly, shows the IPv4 block numbers that have been traded using Section 8.3 of ARIN’s Number Resource Policy Manual. The list contains nothing but IP address block numbers. Using the list of IP Address Blocks Transferred per NRPM 8.3 in combination with ARIN’s Whois and Whowas databases, one can determine the organizations involved, the date of the original registrations and the date of the transfer.

APNIC provides the most information. It maintains a public log of all IPv4 transfers that includes the name of the releasing organization, the name of the acquiring organization and the effective date of the transfer. The problem with the APNIC list is that it includes both market transfers and non-market transfers, and provides no basis for distinguishing between them. Utilization of this data, therefore, involved additional work to see if the organizations involved were part of a merger, or whether they were subsidiaries of the same corporation. In some cases no strong confirmation of the status of a transfer as non-market could be found. In those cases, the transactions were presumed to be market transfers. It is possible that some of those presumptions are wrong, and that we slightly overstate the number of market transfers in the AP region.

Using the methods and data sources described above, we assembled a spreadsheet of all market transfers (including IP numbers transferred as a result of bankruptcy asset sales). Each row contains data showing the region, the IP address of the block traded, the name of the releasing organization, the name of the acquiring organization, the transfer date, the traded block’s prefix, the original registration date of the block, the status of the /8 prefix according to IANA records (legacy or allocated), and the contract type (RSA or LRSA). This data will be made public on the Internet Governance Project web site.

Counting IP Number Trades

In quantifying and analyzing IP number trades, three different units of analysis are relevant: i) the number of transactions, ii) the number of blocks, and iii) the total quantity of IP numbers.

The most obvious metric is the total amount of IP numbers involved. This gives a good overall sense of the dimensions of the trading market. However, for reasons related to routing efficiency, no one sells or buys individual IP numbers; in that respect numbers should not be discussed as if they were homogenous, interchangeable commodities that come in single units.

IP numbers are almost always assigned and allocated in contiguous blocks of various sizes. RIR policies dictate that the smallest block that can be traded is a /24, which consists of 256 unique

---

4 ARIN NRPM, Section 8.3, Transfers to specified recipients. [https://www.arin.net/knowledge/statistics/transfers_8_3.html](https://www.arin.net/knowledge/statistics/transfers_8_3.html)
5 [http://internetgovernance.org](http://internetgovernance.org)
6 Based on technical aspects of routing and network management, one would expect larger contiguous blocks to command higher per-address prices than smaller blocks. Put simply, the larger the block of contiguous numbers a network has, the easier it is for a network operator to manage addressing and routing.
contiguous numbers. The largest allocated block, known as a /8, consists of approximately 16.78 million unique contiguous numbers. Thus, it is also relevant to count and analyze the number of blocks involved in trades, while recognizing that blocks vary widely in size. (Table 4 provides a list that maps the /X notation to specific quantities of IP numbers.)

Finally, one can count transactions. A single market transaction can involve one or more blocks. Generally if the same two organizations are involved in a transfer of multiple blocks on the same date, we count it as one transaction. We found evidence of what may be complicated three-party and four-party transfers of address blocks which may not meet these assumptions, but there are only a few of them.

Number Block Use

Our research also performs some analysis of whether traded address blocks were used before and after the transaction. The lookups were performed using a Python library written by Asghari called PyASN. It takes as input a BGP dump file first. The website Routeviews archives these dumps for many years. Using a script, the IPv4 address space was enumerated, and for each /24 IP block, an ASN lookup is performed using the Routeviews data. We tested at 6-month intervals from January 2010 to July 2012, inclusive.

The Number Market Data

Summary statistics for the period from November 2009 to the end of June 2012 indicate that there were 83 distinct transactions, 204 distinct blocks traded, and a sum total of 6,034,688 unique IP numbers exchanged in the transfer market. Most of the prices are not known, but based on known prices from bankruptcy court records, the overall value of this market is in the tens of millions of dollars at least, and probably around US$ 60 million in total. There was minimal activity in 2009 and 2010, but in 2011 and 2012 the number of transactions, blocks traded and IP numbers began to increase rapidly (Table 2). If the data for the first 6 months of 2012 are extrapolated forward, the number of transactions and IP numbers traded is on pace to increase by a factor of 4 and 10, respectively, from 2011 to 2012, while the number of blocks traded is projected to increase by 60%.

When # IP numbers is used as the metric, North American activity dominates the market, whereas if the metric is # blocks or # transactions the Asia Pacific region is more active. Ninety of the 204 traded blocks making up 5.057 million IP Numbers were in the ARIN region. The # IP numbers traded in the ARIN region is 84% of the total amount traded, whereas ARIN accounts for only 37% of the # blocks. While there seem to be more transactions in the AP region (52 out of 83, or 63%), they generally involve smaller blocks and thus a much smaller quantity of IP numbers in total.

Table 2
Accelerating pace of market transfers (all regions)

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012 1st H</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># transactions</td>
<td>3</td>
<td>2</td>
<td>27</td>
<td>52</td>
<td>83</td>
</tr>
<tr>
<td># blocks traded</td>
<td>8</td>
<td>3</td>
<td>109</td>
<td>84</td>
<td>204</td>
</tr>
<tr>
<td># IP numbers</td>
<td>11,264</td>
<td>10,240</td>
<td>1,013,248</td>
<td>4,999,936</td>
<td>6,034,688</td>
</tr>
</tbody>
</table>

Note that ARIN has not yet run out of IPv4 numbers. At the end of July 2012, ARIN still had 52.6 million numbers available for allocation. The Asia-Pacific region, on the other hand, reached its last /8 on April 15, 2011 and now hands out only small allocations, one to an organization. In other words, North American organizations are turning to the market for IPv4 numbers when they could get numbers from ARIN. Like ARIN, RIPE-NCC still has numbers available; as of July 30, 2012, its website shows 27.38 million in stock. While there have been some parties willing to list blocks for sale in the RIPE region, there have been no buyers and no known transactions in Europe.
Table 3
Quantitative comparison of ARIN allocation and market allocation

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012 1st H</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP numbers allocated by ARIN</td>
<td>41,317,376</td>
<td>45,266,688</td>
<td>22,471,424</td>
<td>16,077,056</td>
</tr>
<tr>
<td># IP numbers allocated by market in ARIN region</td>
<td>11,264</td>
<td>10,240</td>
<td>1,150,976</td>
<td>4,221,184</td>
</tr>
<tr>
<td>% market of total allocations</td>
<td>0.03%</td>
<td>0.02%</td>
<td>5.12%</td>
<td>26.26%</td>
</tr>
</tbody>
</table>

The market constitutes a substantial portion of total allocations in the North American region. Table 3 shows the quantity of IP numbers allocated by ARIN in the normal way from 2009 to the first half of 2012, and compares it to the quantity of IP numbers allocated via market transfers. The quantity of IP numbers involved in market allocations went from 3.6% of administrative allocations in 2011 to 26% in the first half of 2012. If the quantity of numbers involved in market transfers in the ARIN region continues to increase at the pace of the last two years, market transfers could equal administrative allocations in 2013.

Table 4 breaks down market transfers by address block size. It shows that the old Class C (/24) address blocks are the most commonly traded, with 69 transferred blocks, but due to their small size, they account for a tiny portion of the traded numbers. The old Class B (/16) blocks, which contain 65,536 numbers, are also popular objects for trades, with 26 of them changing hands. Twenty-seven /20s were traded as well. Overall, the 35 largest blocks that were traded, from /16 up to /12, account for 91% of all the numbers traded. Still missing from this picture is a /8, the largest unit of allocation. If and when one or more of those blocks trade, all the other trades will shrink to insignificance in relative terms.

Table 4
Address blocks traded

<table>
<thead>
<tr>
<th>/ Notation</th>
<th>Number of blocks this size traded</th>
<th>Numbers per block size</th>
<th>Total number of traded numbers per block size</th>
<th>Percent of all traded numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>/7</td>
<td>0</td>
<td>16,777,216</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>/8</td>
<td>0</td>
<td>8,388,608</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>/9</td>
<td>0</td>
<td>4,194,304</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>/10</td>
<td>0</td>
<td>2,097,152</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>/11</td>
<td>0</td>
<td>1,048,576</td>
<td>2,097,152</td>
<td>33.0%</td>
</tr>
<tr>
<td>/12</td>
<td>1</td>
<td>524,288</td>
<td>524,288</td>
<td>8.2%</td>
</tr>
<tr>
<td>/13</td>
<td>3</td>
<td>262,144</td>
<td>786,432</td>
<td>12.4%</td>
</tr>
<tr>
<td>/14</td>
<td>3</td>
<td>131,072</td>
<td>524,288</td>
<td>8.2%</td>
</tr>
<tr>
<td>/15</td>
<td>5</td>
<td>8,192</td>
<td>114,688</td>
<td>1.8%</td>
</tr>
<tr>
<td>/16</td>
<td>27</td>
<td>4,096</td>
<td>110,592</td>
<td>1.7%</td>
</tr>
<tr>
<td>/17</td>
<td>14</td>
<td>2,048</td>
<td>28,672</td>
<td>0.5%</td>
</tr>
<tr>
<td>/18</td>
<td>20</td>
<td>1,024</td>
<td>20,480</td>
<td>0.3%</td>
</tr>
<tr>
<td>/19</td>
<td>16</td>
<td>512</td>
<td>8,192</td>
<td>0.1%</td>
</tr>
<tr>
<td>/20</td>
<td>69</td>
<td>256</td>
<td>17,664</td>
<td>0.3%</td>
</tr>
<tr>
<td>Sum</td>
<td>204</td>
<td>--</td>
<td>6,034,688</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
The Reallocation of Legacy Blocks

The blocks in play in this market are overwhelmingly comprised of legacy allocations. Of the 6.03 million IP numbers that have been traded, 5.33 million, or 88 percent, were from allocations made before July 1997. 5.36 million, or 89%, were classified as 'legacy' allocations in IANA or RIR records. 100% of the larger blocks sold (/15 and up) were legacy allocations, and 21 of the 26 /16s traded (81%) were legacy blocks. Further, while the releasing organizations are corporations of a highly varied type, the recipients are almost entirely Internet access providers (both mobile and fixed), online service providers such as Amazon and Microsoft, smaller VoIP providers and telephone cooperatives, and hosting companies. This indicates that market transfer policies are succeeding in re-allocating the inefficiently allocated legacy blocks from entities with an unneeded surplus to growing Internet businesses that need them more.

Three of the larger trades of legacy allocations involved bankruptcies in which number blocks were sold to creditors as assets: Nortel, Borders and Teknowledge. The other legacy block transaction involved pharmaceutical company Merck. In 1992 it was given a /8 and from that original allocation it sold two /12s (roughly 2.1 million numbers) to Amazon early in 2012. Borders and Teknowledge sold off legacy /16s as part of their bankruptcy proceeding.

In March 2011, it was announced as part of Nortel’s U.S. bankruptcy proceeding that Microsoft would be acquiring 666,624 IPv4 numbers from Nortel for $7.5 million. Microsoft bought 38 number blocks that had been accumulated since 1989 by Nortel from IANA or from corporate acquisitions. Included in the package were sixteen /24s, four /23s, one /22, two /21s, four /20s, nine /16s, and one /17 and /18 each. A second tranche of Nortel IP numbers, sold as part of the Canadian bankruptcy process, went to Vodafone, Salesforce.com, Bell Aliant, and two smaller ISPs. The Canadian court has refused to release any information about the price of these transactions. The Teknowledge /16 sold for $590,000, or $9.00 per address.

The Merck-Amazon deal was not a bankruptcy but a straight legacy transaction, so we do not know the price. But it illustrates the market’s success at moving IPv4 address stock from legacy holders with excessive allocations to expanding, network-intensive industries. According to our tests, both of the /12 blocks went from being unrouted (i.e., not used on the Internet) to publicly routed within a year of the transaction.

Some Policy Issues

The previous material set out the basic parameters of the transfer market. A solid empirical outline of the transfer market is intended to provide a stronger basis for discussion of the many interesting and important policy questions raised by the future of IP addressing. In this section, we introduce briefly some of those policy issues.

Why ARIN/North America?

One would expect market transfers to take place in APNIC’s region, where the RIR has almost nothing to give out to applicants, and one would not expect to see a lot of market transfers in Europe, where the RIR still has unallocated numbers to give out. Both of those expectations hold up. But North America is the anomaly; ARIN still has numbers to give – more than twice as many as RIPE – and lots of market activity. From a research standpoint, this is an interesting puzzle.

In explaining this puzzle, the Microsoft-Nortel deal is especially revealing. By paying $7.5 million, Microsoft invested about $11.25 per IPv4 address. Using ARIN’s fee schedule for numbers available in its free pool, Microsoft would have paid only $87,250 per year or about 13 cents per address per year in

---


ARIN fees. To pay ARIN $7.5 million in annual fees, Microsoft would have had to hold the address blocks and pay ARIN fees for 86 years, an unlikely eventuality. The disjunction between what Microsoft paid Nortel and what it would have paid ARIN for perfect substitutes indicates that there are factors governing firms' economic calculations regarding IPv4 numbers that may not be obvious to casual observers.

The explanation for this puzzle, we believe, can be found in two policy factors. One is the large gap between the restrictiveness of ARIN's "needs assessment" policies when applied to its remaining free pool allocations and when applied to transfer markets. The other explanation lies in the disjunction between the de facto property rights enjoyed by legacy holders, and the far more limited use rights of non-legacy holders.

**Needs Assessments**

A critical policy issue in the IP number market is the role of administrative needs assessment in the market transfer process. All the transfer policies in place require buyers of number blocks to justify their acquisition by providing technical and business data showing that they "need" the numbers. While market transfers introduce price signals and economic incentives into IPv4 number allocations, the commoditization of the resource is severely limited by the retention of needs assessment. Data from ARIN indicate that in 2011, one third (33%) of all attempts to conduct market transfers via the 8.3 process were blocked or modified due to needs assessments. In 2012 the percentage was 28%. Need assessments thus restrict willing sellers from transacting with willing buyers in about 30% of the cases.

A key variable in the application of needs assessment is the time horizon employed. Proving that one "needs" X numbers tomorrow because one's network is overloaded is a fairly straightforward technical-operational calculation. Proving that one needs X numbers over the next three years in order to accommodate growth and/or implement new business plans is more like an investment decision than a network engineering decision. In free pool allocations, ARIN and the other RIRs have reduced the time horizon for needs assessments as the pool dwindles and exhaustion nears. The time horizon for demonstrating need is now only three months. To qualify for needs assessments in ARIN's 8.3 transfer process, on the other hand, the time horizon for assessing need was one year from 2009 – 2011, and was extended to two years starting in February 2012. Forward-looking companies that want to secure access to IPv4 numbers over a commercially relevant time frame would obviously opt for the transfer market over a free pool allocation – even if the apparent cost of the transferred numbers is much higher.

**Property Rights**

There is another explanation for the existence of a market prior to the depletion of the free pool in North America. Transactions with legacy holders can provide buyers with more secure property rights – although this issue is partially unsettled. The prominent role of legacy holders in the number market has raised important legal and policy questions about the legal rights of the transacting parties. If legacy holders have no binding contract with ARIN, they are not obligated to transfer their IPv4 number holdings via the Section 8.3 transfer policy. This means that legacy holders could transact with buyers regardless of whether the buyers can "demonstrate need." It would also mean that the buyers of legacy numbers would hold them free and clear of ARIN contracts, just as the seller did.

Fearing that such transactions would undermine its authority over a substantial portion of the IPv4 number space, ARIN has agitated to keep transfer market participants within its process. It has even gone so far as to publish advice in a bankruptcy law journal (Ryan & Martel, 2012). As it lacks any contractual leverage over legacy holders, however, it has had to grasp for other forms of influence over legacy sellers and prospective buyers. Specifically, it is attempting to use its control of the Whois database as a strategic lever. ARIN is now warning buyers of legacy resources that it will not update its Whois records to reflect transfers that take place outside of its 8.3 process. It is unclear what the effect of excluding transactions from the Whois will be. It could undermine the value of purchased number blocks, if their absence from the Whois prevented network operators from considering the buyer to be the legitimate holder of the block. This might cause ISPs to refuse to route traffic to the affected number blocks. On the other hand, it is possible that buyers will discount this threat and purchase the block anyway. If the transaction involves a major, reputable corporation and a large block such as a /8 or /16, it seems unlikely that the entire Internet would filter out the number block simply because ARIN didn’t
approve of the trade. If other ISPs routed to the “illicitly” traded block anyway, ARIN’s database would lose its status as an authoritative, reliable guide to who holds which number blocks. Thus, ARIN’s attempt to gain leverage over the transfer market through the use of the Whois database is like a game of chicken; if neither side gives in there could be a collision.

The MSFT-Nortel deal brought these issues to a head. Nortel and Microsoft bypassed ARIN’s 8.3 transfer process and transacted independently to exchange their property rights over the number resources in bankruptcy court. A last-minute intervention by ARIN, and private appeals to Microsoft, led to a compromise solution. Microsoft agreed to sign a special contract for legacy holders, known as a LRSA, and ARIN agreed that the transaction gave Microsoft the same de facto property rights held by the prior legacy holder, Nortel. The specific terms of the LRSA Microsoft signed have not been disclosed.

ARIN claims that it performed a "needs assessment" prior to the Microsoft-Nortel transaction to ensure it was compliant with its policy. Others have disputed ARIN’s claim, dismissing the assessment as a face-saving exercise to make it appear as if it was applying its policy and retaining some authority over legacy address transactions. The evidence supports the more cynical view. By tracing the routing of prefixes, we can see which of the Nortel blocks Microsoft is actually using a year after the trade was approved. We find that as of July 2012, only 7 of the 38 Nortel blocks, totaling only 10,496 numbers, are now being routed by MSFT (Table 5). Indeed, three of the larger /16 blocks transferred went from being routed to being unrouted. Thus, only 18% of the blocks involved in the Nortel - MSFT transaction were routed within a year of the transaction, and due to the withdrawal of the /16s from use there was a net decrease of 186,112 in the quantity of routed IPv4 numbers.

In what sense did Microsoft “need” these IPv4 numbers? From a technical point of view, it clearly did not need to put them into service within the short (one-year) time horizon contemplated by ARIN’s policy. But from a business point of view, it makes perfect sense for an Internet-dependent firm with a market capitalization around $260 billion to spend a paltry $7.5 million to secure long-term access to a resource so critical to its ongoing operations.

Conclusions

There is a thriving and growing market for IPv4 number blocks. Not only does the market exist, but it seems to be doing precisely what its advocates said it would do, namely provide access to additional IPv4 resources after the free pool is depleted, while reallocating number resources more efficiently by moving them out of unused or underutilized allocations and toward organizations who need them to grow.

An additional, partly unintended consequence of the transfer market has been to provide liberalized access to IPv4 numbers relative to the stringent needs assessment and documentation procedures required to get numbers from the RIRs. Companies have shown that they are willing to pay substantially more for IP number resources via the transfer market, if it allows them to extend their time horizon and/or avoid the needs assessment process. Thus the market provides a check on ARIN’s policy process. But due to the confidentiality of the needs assessment and contracting process, it also introduces potential discrimination in the RIR’s contracting process. It may be that there is one set of rules and contracts for smaller, less influential firms and quite another set for larger players whose defection from the ARIN regime might have a large impact. As technical needs assessments become less relevant, there is also the question of how a rising price for numbers will affect competitive entry into the market, and whether price manipulation through hoarding will occur.

It is too early to assess the impact of the transfer market on the migration to IPv6. However, the growth of IPv4 markets occurred simultaneously with what Huston (2012) has characterized as a “fourfold increase in the penetration of IPv6.” Huston’s research recognizes that that rate of IPv6 adoption will not be fast enough to “avoid some of the major pitfalls associated with encountering IPv4 exhaustion.” In other words, here again the market seems to be doing exactly what its advocates said it would do, namely provide a bridge over a period of scarcity that is of indefinite duration while Internet operators gradually come to terms with the costs and technical issues associated with IPv6 implementation, and wait for the great tipping point. To conclude, it is time to stop debating the merits of IPv4 address markets; the issue now is how to make them work better.
Table 5
Utilization of Nortel blocks by Microsoft

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>131.253.1/24</td>
<td>ASN 19952 Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.3/24</td>
<td>ASN 19952 Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.5/24</td>
<td>ASN 19952 Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.6/24</td>
<td>ASN 19952 Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.8/24</td>
<td>ASN 19952 Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.12/22</td>
<td>ASN 19952 Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>ASN 8075</td>
<td>ASN 8075</td>
</tr>
<tr>
<td>131.253.16/23</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>ASN 23468</td>
</tr>
<tr>
<td>131.253.18/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>ASN 8075</td>
</tr>
<tr>
<td>131.253.21/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.22/23</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.24/21</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.32/20</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.61/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.62/23</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.64/18</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>131.253.128/17</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>132.245.0/16</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>134.170.0/16</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>134.177.0/16</td>
<td>ASN 7099</td>
<td>ASN 7099</td>
<td>ASN 7099</td>
<td>ASN 7099</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>137.116.0/16</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>137.117.0/16</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>137.135.0/16</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>138.91.0/16</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>141.251.0/16</td>
<td>Unrouted</td>
<td>ASN 7099</td>
<td>ASN 7099</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>192.32.0/16</td>
<td>ASN 7099</td>
<td>ASN 7099</td>
<td>ASN 7099</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>192.48.225/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>192.84.159/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>192.84.160/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>192.84.161/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>198.49.8/0/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>ASN 8075</td>
</tr>
<tr>
<td>198.200.130/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>198.206.164/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>199.30.16/20</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>199.74.210/24</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>ASN 8075</td>
</tr>
<tr>
<td>199.24.32/20</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>199.24.48/21</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>ASN 8075</td>
</tr>
<tr>
<td>204.152.140/23</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
<tr>
<td>205.174.224/20</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
<td>Unrouted</td>
</tr>
</tbody>
</table>
References


An Exploratory Study of Users’ Preferences and Use of Sources for Seeking Health Information

Yan Zhang
School of Information
University of Texas at Austin
yanz@ischool.utexas.edu

Abstract

Studies have surveyed college students’ use of health information sources. But still little is known about the cognitive processes involved in their source evaluation and selection. To explore this topic, we conducted in-depth interviews with twenty students about their use of sources for specific health-related incidents in their lives. The results suggested that search engines, health websites, doctors, and parents were the most important sources, while social media and libraries were the least important ones. The participants cared about the quality and accessibility of information, but also inclined to use sources that they are familiar and comfortable with, sources that match their interests, or sources that are personal and engaging. In actual information seeking, some participants relied only on search engines, while more used multiple sources and involved family, friends, or doctors in the process. The results’ implications to the design of health websites and health education programs were discussed.

Keywords: consumer health informatics, information behavior, information sources, Web 2.0, social media

Introduction

According to the American College Health Association’s (ACHA) most recent National College Health Assessment (NCHA), 92.7% of the students sampled reported their health as being “good”, “very good”, or “excellent”. Nevertheless, 55.3% of students also reported being treated or diagnosed for a specific disease or condition by a health professional in the last 12 months and 25.0% had some form of a chronic condition or disability (American College Health, 2011). The survey, conducted by the Pew Research center, also reported that 71% of young adults (aged 18-29) look online for health information, the highest among all generations (Fox, 2011). Furthermore, searching for health information is the third most popular online activity for this group, after email and search engine use (Zickuhr, 2010). These results suggest that college students, although generally healthy, have a high demand for health information, and, are diligent in seeking health information.

Purposeful information seeking always begins by selecting sources (Wilson, 1999). Traditionally, college students refer to various sources for health information, such as healthcare professionals, family and friends, TV, newspapers, magazines, and the Internet (Kwan, Arbour-Nicitopoulos, Lowe, Taman, & Faulkner, 2010). In recent years, the rapid development of Web 2.0 and mobile technologies led to the emergence of an array of new forms of sources, such as online communities, social Q&A sites, and social networking sites, where consumers share information and personal stories and construct knowledge (Eysenbach, Powell, Englesakis, Rizo, & Stern, 2004). As a population at the forefront of embracing and adopting new technologies, college students’ information environment for health has been significantly expanded; nevertheless, challenges are present when trying to use and interpret information in these new media (Kim & Sin, 2011). In any context, source selection can have an important effect on the success of an information searching attempt; this effect can have severe ramifications when the topic is related to personal health information. Thus, it is important to understand, in the current more complicated information environment, what sources college students use and prefer, and how they evaluate the sources.

Acknowledgements: This study was supported by the Alumni Fellowship from the School of Information at the University of Texas at Austin. The author wants to thank the participants for their contributions.


Copyright is held by the author.
Studies have been conducted to explore this topic; however, many were self-administered surveys and, often, only focused on a particular source (such as the Internet) (for e.g., Escoffery, et al., 2005; Kwan, et al., 2010). Such studies shed light on students’ preferences of sources, but are limited in revealing how they actually use and evaluate these sources. Furthermore, few studies examined how students view the emerging new media as sources for health information. In order to achieve a more comprehensive understanding of college students’ source preferences and use in this increasingly more diverse and dynamic information environment, we adopted the in-depth interview technique to reveal criteria used by them to select and use different sources to address a particular issue of interest. Specifically, the following research questions were explored:

1. What sources do college students use when looking for health information?
2. What criteria do they use to select sources?
3. How do they use the sources to address specific health situations?

**Background**

Selection of proper information sources precedes successful information seeking. Thus, source preference and use has been an important research topic, the understanding of which is deemed important for understanding information seeking behavior as well as designing effective health promotion programs and health websites. To understand people’s preferences and use of information sources, several conceptual frameworks were developed in both Information Science and Communication Studies, such as information field (Johnson, Case, Andrews, Allard, & Johnson, 2006), information horizons (Sonnenwald & Iivonen, 1999), and information zones (Savolainen, 2008). A common assumption of these frameworks is that everyone is situated in a virtual information environment. This environment encapsulates an array of sources, such as information retrieval tools, the Web, books, magazines, and social networks, such as family, friends, and colleagues (Sonnenwald, Wildemuth, & Harmon, 2001). These sources cluster into different zones, dependent on their importance to the users (Savolainen, 2008). Furthermore this environment serves as a starting point for information seeking and constitutes part of the context that fosters or constrains people’s access of information (Wilson, 1981).

College students are active consumers of health information (Fox, 2011). For this group of users, major topics of interest included fitness/exercise and diet/nutrition, life-style choices such as alcohol, smoking, and illegal drug use, as well as sexual health, such as STDs, contraception, and dating violence (Baxter, Egbert, & Ho, 2008; Buhi, Daley, Fuhrmann, & Smith, 2009; Escoffery, et al., 2005; Hanauer, Dibble, Fortin, & Col, 2004). Students were also interested in topics such as cancer, diabetes, and mental health, especially when family histories indicated a high risk for these diseases (Escoffery, et al., 2005).

Numerous studies have examined where college students receive health information. It is widely recognized that, in this digital age, the Web has become a leading source for this group. The percentage of students who used the Web for health information increased from 73% in 2005 to nearly 79% in 2010 (Escoffery, et al., 2005; Kwan, et al., 2010; Percheski & Hargittai, 2011). In most studies, the Web was treated as one entity and few attempted to examine roles that different platforms, such as search engines, health websites, and social networking sites, play in students’ health information searching (Buhi, et al., 2009). Although being the most prevalent source, the Web is not the most frequently used one, as students had doubts about the credibility of information on the Web (Escoffery, et al., 2005). Human sources seemed to be a preferred source. Studies consistently reported that family and friends were the most frequently used sources, followed by medical professionals, such as doctors and nurses (Percheski & Hargittai, 2011; Siebert, Wilke, Delva, Smith, & Howell, 2003). These sources were also cited as more credible than the Web (Vader, Walters, Roudsari, & Nguyen, 2011). Other human sources included health educators, romantic partners or spouses, and acquaintances (Baxter, et al., 2008). Although, in many domains, the Web has challenged the use of traditional mass media, such as TV, radios, and magazines, these sources are still used by more than half of the students for health information (Kwan, et al., 2010; Percheski & Hargittai, 2011).

To gain an in-depth understanding of students’ source preferences and use, it is necessary to understand how they evaluate the sources. Nevertheless, most existing studies focused on examining criteria that cross-sectional consumers employ to evaluate health information on the Internet. These studies consistently suggested that consumers evaluate health information by appraising both the quality and characteristics of information, such as authoritativeness, trustworthiness, currency, and readability, as well as the design attributes of the websites, such as system usability, interface appearance, and...
information organization and presentation (Cline & Haynes, 2001; Eysenbach & Köhler, 2002; Morahan-Martin, 2004). Based on observations of fifteen menopausal women searching for information, Silence, Briggs, Fishwick, & Harris (2004) further proposed a staged model of online health information evaluation. At the first stage, users quickly reject certain websites, mainly dependent on design factors (e.g., layout, navigation aids, and interactive features); at the second stage, users meticulously select websites to seek information. This selection is mainly dependent on the appraisal of content factors (e.g., accuracy, topic coverage, and readability).

Although the results from the reviewed studies were not specific to college students, they may also apply to this group. For example, in one study that we were able to found, Escoffery et al. (2005) surveyed 743 college students on their use of Internet for health information and found that most of them ranked attributes associated with information, specifically accuracy, credibility, currency, comprehensiveness, easy of understanding, and readability, as important factors for judging a health website. The usability of the sites was also an important factor, while website appearance, use of multimedia, and interactivity were reported as less important. In two additional surveys, students were simply asked to rank the believability of a set of health information sources (Kwan, et al., 2010; Vader, et al., 2011). Another survey-study reported that many students lacked the ability to judge the trustworthiness of health-related websites and articles (Ivanitskaya, O’Boyle, & Casey, 2006).

Based on the review, it is apparent that most existing studies of students’ evaluation of health information sources relied on survey instruments. This method has limitations. First, the imposed nature of surveys limits participants’ answers to the criteria provided and it is difficult to capture their interpretations of each criterion. Second, limited information about the context in which the criteria were used can be captured by surveys. Thus, in this study, we adopted the in-depth interview technique, intending to understand, from the students’ perspective, what criteria they use to evaluate and select sources and how they used the sources in specific incidents. Furthermore, most existing studies focus on students’ evaluation of health websites. In this study, we extend our exploration to include other sources students are exposed to, including human sources, traditional mass media channels, social media technologies, government social services, and libraries. A more comprehensive and in-depth understanding of students’ preferences and use of various sources is needed as their health information environment becomes unprecedentedly diverse and complex.

Methods

Twenty undergraduate students from a major university in Texas were interviewed about their use of information sources for seeking health and wellness-related information. These participants were recruited through an email message sent to a campus-wide mailing list. A screening survey was distributed to those who expressed interest in the study to collect demographic information and experience related to searching for health information. The screening was in place to ensure that participants had seriously searched for health information for themselves or for people they cared about. We selected the first twenty participants who met the criteria.

The interviews were one-on-one and took place in a private lab or office. The interviews were conducted from January to April of 2012. At the beginning of each session, a brief introduction of the study was given and the participant was then asked to review the consent form. After giving the consent, each participant was presented with a list containing sixteen health information sources, including family members, friends, someone else with similar conditions, doctors or other healthcare professionals, government agencies, web search engines, health websites, and Web 2.0 sources, such as Wikipedia, blogs, social networking sites, and YouTube. The list was generated based on a review of the literature on consumer health information sources (for e.g., Fox, 2011; Kwan, et al., 2010), as well as our interest in students’ use of Web 2.0 sources for health information. Each participant was asked to select sources that they had used for health information. Then he/she was asked to rank each source on a 0-3 scale in terms of their perceived importance of the source for their health information seeking: 3- very important (primary), 2- important (secondary), 1- less important (tertiary), and 0-not important (not-used). The interviewer also asked the participant to describe his/her use of each selected source and explain the ranking assigned to the source. Then, the participant was asked to recall a most recent or most memorable experience of seeking health information and describe the information seeking process, specifically, sources used, the reasons for choosing the sources, and strategies of using the sources. At
the end of the interview, the participant was asked to draw a timeline view of the sources that he/she had used for this incident. Each interview lasted approximately 60-90 minutes.

Users’ demographic information, sources used, and sources ratings were analyzed using descriptive statistics. The interviews were transcribed and analyzed using a qualitative content analysis method by following an open coding process (Glaser & Strauss, 1967). The coding unit was a theme, specifically the criterion that the participants used to select and evaluate each source. The data analysis was assist with Nvivo software. A second coder coded 50% of the transcripts and the percentage agreement reached 89.2%.

Results

Characteristics of the Participants and Health Issues of Interest

Of the twenty participants, twelve were females and eight were males. Their ages ranged from 18-24 years old (Mean = 19.8; S.D. = 2.0). Their experience with searching for health information ranged from 1-10 years (Mean = 3.94; S.D. = 2.4).

The topic of interest helps shape users’ preferences and use of sources (Savolainen, 2008). To place college students’ source preferences into context, we asked the participants to describe a specific health information seeking incident. They were also encouraged to talk about the topics that they searched for when they ranked and described the use of the sources. The topics mentioned by the participants roughly fell into the following categories:

- Exercises, fitness, and physical training, such as workout tips and methods; use of supplement (e.g., creatine) for powerlifting.
- Healthy food and diets, such as calorie and nutrition of food (e.g., salad, fruits, and vegetables).
- Weight loss and weight control
- Physical injuries, such as ankle injury, Morton's neuroma, back injury, knee injury and treatments of the injuries, such as physical therapy
- A particular health concern (symptoms or conditions), for example, do I have ear infection? What are the bumps around the wrist and knuckles that itched a lot? What are the large rashes along my thighs and lower back? Why am I sick after returning from a vacation? Why do I have unusual menstrual bleeding?
- Mental health, such as stress and depression
- Preventative health, for example, how to protect myself from contracting swine flu and when to take seasonal flu shots

Similar to previous findings, the most popular topics of interest to the participants were exercise and healthy diets, with more than half reporting having searched for these topics. The results also reflected that undergraduates had a wide range of other health-related requests and concerns, as demonstrated in the list. One participant also mentioned searching for information online and discussing diabetes with family members because of a family history of the disease. It seemed participants’ searches for health information were motivated either by a personal interest (e.g., exercises, diets, and diabetes) or by a need to solve a particular health concern (e.g., what are these rashes?).

Sources Used for Health-Related Information

The participants’ health information horizon varied. Among the 16 sources provided, each participant had used eight to twelve (Mean = 10.2; SD = 1.5). Table 1 lists the participants’ ranking of each source in terms of their evaluation of the importance of the source for their health information seeking. The 16 sources were grouped into 6 major categories.
Table 1
Sources used for health-related information, ranked by the perceived importance

<table>
<thead>
<tr>
<th>Sources</th>
<th>Mean</th>
<th>Primary</th>
<th>Secondary</th>
<th>Tertiary</th>
<th>Not use</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traditional Internet sources</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web search engines</td>
<td>2.60&lt;sup&gt;a&lt;/sup&gt;</td>
<td>14&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Health websites</td>
<td>2.45</td>
<td>14</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Doctors or other healthcare providers</td>
<td>2.45</td>
<td>13</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Human resources</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family Members</td>
<td>2.25</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Someone else with similar conditions</td>
<td>2.15</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Friends</td>
<td>1.85</td>
<td>3</td>
<td>11</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><strong>Traditional media</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Books/newspapers/magazines/pamphlets/other printed material</td>
<td>1.55</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Television</td>
<td>.90</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1.60</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Blogs</td>
<td>.90</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td><strong>Web 2.0 sources</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support groups/online communities</td>
<td>.45</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Generic social networking sites (e.g. Twitter, Facebook)</td>
<td>.20</td>
<td></td>
<td>4</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>RSS feeds/email newsletters</td>
<td>.45</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>YouTube</td>
<td>.60</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td><strong>Government services</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government agencies/social services/health centers</td>
<td>.60</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td><strong>Libraries</strong></td>
<td>.45</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>15</td>
</tr>
</tbody>
</table>

<sup>a</sup> 3= very important (primary), 2= important (secondary), 1= less important (tertiary), and 0=not important (not-use)
<sup>b</sup> Number of participants

As shown in the table, the participants ranked traditional Internet sources, namely, Web search engines and health websites, as highly important sources for health information. The most widely referred to Web search engine was Google; only one participant mentioned Yahoo! and one mentioned Bing. The health websites mentioned by all of the participants was WebMD; others included the CDC website, Mayo Clinic, the FDA website, Yahoo! portal, and Dr. Oz’s website. Consistent with findings from many previous studies, doctors and other healthcare providers were ranked as one of the most important sources for health information. The participants’ social ties, particularly family members and someone with similar conditions, mostly friends, were also deemed important. Although traditional medial sources, namely printed materials (books, newspapers, magazines, and pamphlets) and mass media (TV), received a lower ranking, they were still favored. Specific print materials mentioned by participants included Cosmo, Vogue, Women’s Health magazine and pamphlets in doctors’ office; TV programs mentioned included CNN, Dr. Oz, and the Martha Stewart Show.

College students are early adopters of various Web 2.0 technologies. Nevertheless, when it comes to look for health information, Web 2.0 sources were not favored and were also underused. Among various forms of social media, Wikipedia was considered as the most important source for health information. The remaining sources, including blogs, online communities, social networking sites, and RSS feeds, were used by less than half of the participants, and few considered them as important sources. Those who used these new media mentioned the following specific sites: bodybuilding community, Kimberly Snyder’s blog, and Yahoo! Answers. Only a few participants had used government-run social services; subsequently, this source was ranked low in its importance. Libraries were ranked as the least important source on the list, with 75% of the participants not using libraries for health-related information.

In addition to selecting and ranking the sources used from the list, participants were asked whether they had used sources not on the list. Close to half mentioned the University Health Service (UHS) as a source and one mentioned nutrition labels as a source of calorie and nutrition information.
Criteria for Source Selection and Use

To understand why users select and use particular sources over others, we asked the participants to describe reasons for their source selection decisions. The participants mentioned altogether fourteen unique criteria used in selecting and using sources for health-related information. These criteria were collapsed into six major categories pertaining to: the characteristics of content of a source, the characteristics of the source itself, the familiarity of the source to the user, the user’s personal interest, the user’s emotional reactions to the source, and social norms. The six categories can be further classified into two sets: the first two categories are related to the characteristics of sources and source content, while the rest were related to the characteristics of the user.

Set 1: Criteria related to the characteristics of sources and source content. Among all the criteria, the majority (10 out of 14) were related to the perceived quality and characteristics of the sources and source content, as shown in Table 2a.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Criteria</th>
<th>Description</th>
<th>No. of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source content characteristics</td>
<td>Authoritativeness</td>
<td>Whether the content of a source is truthful or correct (when it is human source, it refers to the perceived authority of the human)</td>
<td>18 (90.0%)</td>
</tr>
<tr>
<td>Source content characteristics</td>
<td>Visual presentations</td>
<td>Whether the content is presented in a visual format</td>
<td>10 (50.0%)</td>
</tr>
<tr>
<td>Source content characteristics</td>
<td>Specificity</td>
<td>Whether the content of a source is specific or general</td>
<td>5 (25.0%)</td>
</tr>
<tr>
<td>Source content characteristics</td>
<td>Easy to understand</td>
<td>Whether the content is easy to understand</td>
<td>5 (25.0%)</td>
</tr>
<tr>
<td>Source content characteristics</td>
<td>Currency</td>
<td>Whether the content is up-to-date</td>
<td>2 (10.0%)</td>
</tr>
<tr>
<td>Source characteristics</td>
<td>Accessibility</td>
<td>Whether a source is available or easy to access</td>
<td>20 (100%)</td>
</tr>
<tr>
<td>Source characteristics</td>
<td>Trustworthiness</td>
<td>Whether a source is integral in terms of motivation to be truthful</td>
<td>16 (80.0%)</td>
</tr>
<tr>
<td>Source characteristics</td>
<td>Usability</td>
<td>Whether a source, including the technological platform, is easy to use</td>
<td>16 (83.3%)</td>
</tr>
<tr>
<td>Source characteristics</td>
<td>Similarity</td>
<td>Whether a source (mostly human) is similar to the user in terms of health condition and interest</td>
<td>14 (70.0%)</td>
</tr>
<tr>
<td>Source characteristics</td>
<td>Scope</td>
<td>Whether a source is health information specific or includes many other topics</td>
<td>10 (50.0%)</td>
</tr>
</tbody>
</table>

Source accessibility and the authoritativeness of the content were the most widely used criteria, followed by trustworthiness and usability. Sources were more accessible when they were always available (e.g., search engines and online communities), within reasonable physical distance (e.g., libraries and parents), and can be reached within an acceptable time frame (e.g., you have to make an appointments to see a doctor; friends were with me at the moment). Authoritativeness involves the judgment of the truthfulness and accuracy of the information (O’Keefe, 2002). Sources were authoritative when perceived as having reliable medical knowledge (e.g., doctors, sometimes parents and friends, and Mayo clinic websites) or provided references to other sources for validation (e.g., Wikipedia). Trustworthiness refers to judgments concerning the motivation of a source to be truthful or biased (O’Keefe, 2002). Sources were trustworthy when they were perceived as being genuinely interested in one’s health (e.g., family and friends) and lacked monetary incentives (e.g., TV programs). Usability refers to whether a source is easy to use. Here are several examples:
“Because I mean, I’m not oh where I used to live and didn’t have a library. So I couldn’t really just walk to one. I don’t think they’ve got any health or fitness related type in the libraries nearby on campus.” – [Accessibility]

"[Wikipedia], a lot of people put on their knowledge, there are tons of references. I would say it is fairly valid because most people don’t like, you know, edit it just for the heck of it. There is always like outside links to science, articles you can read. I would consider it good quality." – [The authoritativeness was established by providing references to other publications for validation]

“I used to watch [health-related TV programs], but there were so many. And I was thinking like the main point of advertising is to sell their products so can I trust them? I mean, because they're very biased towards their product. They want to sell their products so I don't think they're a very reliable source.” – [Trustworthiness: TV ads are motivated by profit gains]

“I feel like a library will be a lot harder to find the information I am looking for in entire books but I feel like in WebMD it could probably easier, it more easily gets the information that you need and probably quicker.” – [WebMD is better than a library in its usability]

The other source and source content-related criteria used by more than half of the participants included a source’s (mainly human) similarity to the user in terms of body shape, health conditions, or health interests; visual presentations of the content (mainly in videos and images), and the scope of the source. The following quotes illustrate these criteria:

“Well like, I would listen to like, pertaining to exercise and what to eat and how to like, like build myself better. I do listen to people that are like somewhat the same body type as me.” – [Similarity]

“If something interests me say like cancer I will look it up on Wikipedia and read all about it. I like images and everything.” – [Visual presentation]

“Type in a few keywords [in Google] and websites are coming up. That one looks good. [It is] so much easier to use, lots of diversity, [I can] compare a lot of stuff.” – [Scope]

The rest of the criteria in this group were mentioned by only several participants, including the specificity of information (e.g., “WebMD is more specific [than Wikipedia]”) and whether the information is easy to understand (“I see what I don’t really understand I go to Wikipedia for an easier description”). Only two participants expressed concerns about information currency by pointing out that he/she used online communities for health information because they are “more updated.”

**Set 2: Criteria related to the user: Familiarity, personal interest, affection, and social norms.** The second set of criteria are not related to the features of the sources; rather they are related to the user’s characteristics, including their familiarity to the source, personal interest, affections or emotions involved, and perceived social norms, as shown in Table 2b.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Criteria</th>
<th>Description</th>
<th>No. of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Familiarity</strong></td>
<td>Familiarity/Previous experience/Habits</td>
<td>Whether a source is familiar to the user</td>
<td>15 (75.0%)</td>
</tr>
<tr>
<td><strong>Personal interest</strong></td>
<td></td>
<td>Whether the content of a source is of interest to the user, particularly whether it is engaging or entertaining</td>
<td>15 (75.0%)</td>
</tr>
<tr>
<td><strong>Affections</strong></td>
<td></td>
<td>The selection of a source is dependent on the emotional reactions of the user.</td>
<td>6 (30.0%)</td>
</tr>
<tr>
<td><strong>Social norms</strong></td>
<td></td>
<td>Whether use of a source complies with generally accepted social norms</td>
<td>3 (15.0%)</td>
</tr>
</tbody>
</table>
Familiarity of a source was mentioned by fifteen participants as a determinant in their selection and use of a source. Indicators of familiarity included whether the user was aware of the availability of health information in the source, whether he/she had previous experience with the source, and whether he/she had a habit in using the source (e.g., “used to Google”). Here are two examples:

“I honestly did not know [blogs, support groups, and online communities] existed for health related material.” – [Lack of familiarity to these sources]
I use those sources [search engines] for pretty much, I use them every day pretty much. And I use them for school. So, I think I know how to work them. It’s easier for me. – [Familiarity of Web search engines]

Additionally, fifteen participants pointed out personal interest as a factor in guiding their source consumption, specifically, whether he/she was interested in the content of the source and whether the presentation format was engaging or entertaining. This criterion was mostly used to determine whether the user was going to pay attention to sources in their current environment, such as TV programs, magazines, newspapers, pamphlets in doctors' offices, blogs, and Facebook, and where users often do not actively seek health information, but passively monitor the content. Here is an example:

“Like Dr.Oz, if I'm home in the afternoon and just working on my homework, and he comes on and says eat chAsE berries because they'll do all this for you. Then I'm like that's interesting, or if he comes on with something that I'm like though of. I have dealt with that in the past. What does he have to say about it? Or if it's new research that has been found out.” – [Whether this TV program matches with my personal interest]

Six participants mentioned affections as a factor that impacts their use of a source. One participant commented that he/she did not tell her mom certain health concerns because she did not want her to "get worried." One felt "comfortable" talking to his/her friends and another felt "comfortable" talking with people with similar conditions about very private health concerns on an online forum. Furthermore, one participant expressed comfortableness in talking with a doctor in the University Health Service because she listened and was “genuine” and “concerned.” Three participants cited social norms as a factor preventing them from using social networking sites, particularly Facebook and Twitter for health information; one participant commented: “I don’t go to those websites for health questions. It's too public.”

It is worth noting that the selection of a source is not always a result of the consideration of one criterion. Sometimes, it involves deliberate weighting of multiple criteria. For example, authoritative sources might be difficult to access (e.g., visiting doctors requires appointments); authoritative sources might not fit one’s personal interest (e.g., not comfortable with the doctor in the UHS); sources with an appropriate level of specificity might not be easy to understand (e.g., WebMD content is specific, but not as easy to understand as it is in Wikipedia). The following quote illustrates a case of source evaluation where both accessibility and accuracy of information were weighted:

“When you Google search you end up at blogs sometimes. So, I looked at some of those. But, I guess accuracy is a little bit of concern. Like I don’t trust it as much, I guess, as I would some of the other sources.”

Use of Sources in Seeking Information for Specific Health Situations: A Process View

The criteria identified in the previous section helped illuminate students’ cognitive appraisal of health information sources, but shed limited light on the use of source as a process. Source selection is an integral part of information seeking, thus is a dynamic process. To understand how sources are actually being used, pathways, that is, the sequence in which sources were being used to address specific health situations were analyzed based on the critical incident interviews. Two major patterns were identified: (1) Search engine-centered information seeking, and (2) Multi-source information seeking. Figure 1a and 1b show an example of each pattern. The figures were produced based on the participants’ drawing of their process of searching information for a particular incident.
Figure 1a. Search engine centered information seeking. Figure 1b. Multi-source information seeking.

Figure 1a illustrates the process that a participant followed to search for information about unusual menstrual bleeding for his girlfriend. He began with Google by typing in keywords and then examined a few results one by one. Google was chosen because he was used to the site and the health matter of concern was very personal. He stayed with Google because he believes that all results returned by Google were relevant, as he commented “that’s what Google does, it presents articles that were all relevant to what your keywords are. So, I didn’t want to stray away from that.” Figure 1b demonstrate a multi-source pattern for seeking health information. This participant was trying to find out whether the symptoms that he/she had signified a STDs. He/she began with a Google search (because of easy access and because using Google is a habit), moved onto making a phone call to a call center (gave it up because he/she felt it was commercially motivated), had a brief conversation with a very close friend to hear about his/her opinions, followed by another around of online search, and then went to his/her father, who was a doctor. This multi-source pattern has also been noted in earlier research (Pescosolido, 1992).

Three participants demonstrated the first pattern while the remaining demonstrated the second pattern (with varied number of transitions between varied numbers of sources). Among the participants, eight began their search from a human source, respectively, parents, cousin (doctor), coaches, and friends; seven began from Google; and four began from WebMD. Google was the source that used by all the participants and it was always used as the first or second source in the search process. Additionally, fifteen participants (75%) involved various human sources (father, mom, cousin, coaches, friends, and doctors) at different stages in their information seeking process. It seemed that the participants’ actual use of the sources, particularly search engines and human sources, matched with their ratings of the importance of the sources. Nevertheless, the sequence in which sources were deployed and the reasons why users move from one source to another requires further exploration.

Discussion and Conclusions

This study examined how undergraduate students evaluate and use various sources for health information. The findings suggested that students preferred the Web, particularly search engines and health websites, and human sources for health information, which is consistent with previous research (Case, Johnson, Andrews, Allard, & Kelly, 2004; Fox, 2011; Kwan, et al., 2010; Pennbridge, Moya, & Rodrigues, 1999). Nevertheless, this study contributes to the understanding of college students’ health information seeking behavior by revealing how they use and perceive Web 2.0 sources for health information. The results suggest that, although embracing social media technologies in general, college students are not receptive to using these technologies for health purposes (Fox, 2011). Many were not even aware that these sources contained health information. For those who were, they were concerned about the quality of information. Among these sources, Wikipedia seems to be an exception. Students used Wikipedia because it was often ranked high in Google search results, they trusted the accuracy of
the information resulting from the collective edit process, and the information was easy to understand. In another study, Wikipedia was also reported by college students as a credible source for sexual health information (Buhi et al., 2009). Only a couple of participants expressed a favorable view of social media, pointing out that information in social media is personally relevant and up-to-date.

The second contribution of this study is that, in addition to corroborating previous research on the significant impact of the authoritativeness, accessibility, trustworthiness, and usability of a source on students’ evaluation of health information sources (e.g., Cline & Haynes, 2001), it identified two additional source-related factors: sources’ similarity to users (particularly human sources), as well as, the visual presentation of information. Moreover, it identified a set of criteria related to users’ characteristics, including familiarity with the source, personal interests, affections, and social norms.

The identification of these criteria suggests the following implications for health website design and health education: (1) health websites or social media sites should support users to find people similar to themselves or with similar interests, and be able to accommodate users’ personal health interests. For generic social networking sites, like Facebook, it should allow friends with similar health interests to form groups and communicate privately so as to create a safe and comfortable environment for health information exchange; (2) websites should effectively integrate image and videos into their collections to tailor to users who need procedural information on how to do something (e.g., workout) or to those who are visual learners; (3) to promote use of credible health information sources; it is necessary to increase students’ familiarity with these sources. Universities, university health services, as well as libraries should take up a role as health educators to improve students’ health literacy. Methods like creating health literacy courses in universities, and providing health information search guidelines in libraries could be used. Universities health services could expand their influence by participating in social media and become more connected with students; (4) public health campaigns on mass media should continue to be strengthened to impose positive influence on students’ health and health behaviors, because, as the results suggested, college students also passively monitor their information fields, particularly TV, magazine, newspapers, and Facebook, to receive information of personal interest; (5) health website design should not only be user-centered, but also take users’ emotions and affections into consideration (Norman, 2003), as it is necessary to create an engaging and comfortable virtual environment for users.

Most existing studies on source selection and use in health information seeking utilized the survey method. An inherent weakness of this method is that it examines source selection out of the context of information seeking as a process and operationalize it as one-time decision-making. Thus, the third contribution of this study is that it provided a pathway/process view of students’ use of sources in real information seeking scenarios. Two source use patterns, a search-engine-centered pattern and a multi-source pattern, were identified. The former seems to be a reflection of the least-effort principle, which asserts that people weigh the accessibility of a source over its quality (Anderson, Glassman, McAfee, & Pinelli, 2001); while the latter seems to reflect a collective influence of the cost-benefit analysis, which weighs source quality over accessibility (Ashford, 1986), and the least-effort principle, because in the second pattern, participants weighed both source quality and accessibility (Agarwal, Xu, & Poo, 2011). In future studies, it is worthwhile to examine why users select multiple sources and why they move from one source to the other.

This study is exploratory. In future studies, we will also add the University Health Services (UHS) to the source list, as close to half of the participants mentioned it in the study. Previous research suggested that the criteria students used for selecting sources could be influenced by the topic of interest. For example, potentially embarrassing topics such as sexually transmitted diseases (STDs), contraception, and pregnancy are cited as reasons for seeking information online where anonymity can be maintained (Gray, Klein, Cantrill, & Noyce, 2002). Therefore, in future studies, it is worthwhile to examine the impact of topics on students’ evaluation and use of various sources, which could effectively inform both health education and health website design.
References


Gender as a Predictive Factor for Tasks Completed Using Smartphones

Mark Evans
University of North Texas
markevans@my.unt.edu

Susan Hopper
University of North Texas
susanhopper@my.unt

Greg Jones
University of North Texas
Greg.Jones@unt.edu

Gerald Knezek
University of North Texas
Gerald.Knezek@unt.edu

Abstract

Smartphone usage among undergraduate students is growing (Woodcock, Middleton, & Nortcliffe, 2012). The prevalence of these devices in every aspect of student routines presents the research problem: how will smartphone usage affect how college students manage their coursework? To address this research problem, we conducted a pilot study examining the current level of smartphone usage among college students. A distinct battery of questions were created and posted to an online survey. Forty-two undergraduate students who own and use smartphones participated. Survey questions focused on how frequently students utilized specific smartphone functions to complete class-related tasks. This paper applies a quantitative analysis to this specific battery of questions in an effort to address the research question: Is gender a significant factor in determining the frequency with which smartphones are used to complete informal learning tasks?

Keywords: gender, smartphone, mobile learning, task, behavior

Introduction

Smartphone usage among undergraduate students is growing exponentially (Woodcock et al., 2012). The prevalence of these devices in every aspect of student routines presents researchers with the question: how will smartphone usage affect how college students manage their college coursework and task responsibilities? Initially, this research study sought to address this question through a qualitative analysis of current student interviews. We hypothesized that the level of adoption and how frequently students use their smartphones to manage coursework could indicate the level of impact that smartphones have in a college setting. For this purpose, we took the preliminary step of developing a battery of questions based on observations made while teaching a college-level blended learning class. The initial draft of questions focused on how frequently the students utilized their smartphones to complete specific tasks related to managing their blended-learning coursework inside the classroom. Next, we broadened the scope of the questions to include tasks that could be conducted using a smartphone while outside of the classroom. These were paired down to a final battery of 15 questions testing task frequency for each student, inside and outside of the classroom. Forty-two undergraduate students, 15 males and 27 females, who owned smartphones and were enrolled in a blended learning course at the time of the survey, participated.

After compiling the results, we tested the online survey as a statistical instrument to see if more might be gleaned from the data collected. We hypothesized that a student’s gender could be a key determinant in how frequently students choose to utilize their smartphones to complete different types of


Copyright is held by the authors.
activities relevant to their coursework. This study applies a quantitative analysis to the battery of questions in the online survey, in an effort to test the research question:

Is gender a significant factor in determining the frequency with which smartphones are used for informal learning tasks?

The outcome of this analysis will provide insight into whether a student’s gender might be a predictor of how frequently students use smartphones to complete specific tasks related to managing their coursework.

**Literature Review**

It has been well-established that gender plays a critical role in the formulation of an individual’s personality traits (Bouchard Jr. & Loehlin, 2001). Studies also indicate that specific personality traits have a significant impact on smartphone usage (Lane & Manner, 2011). Clough, Jones, McAndrew, and Scanlon (2008) suggest that smartphones are already, “used extensively in an informal learning context by enthusiasts.” Furthermore, extensive research shows a significant difference in how men and women use social networking sites and web-based media, one of the most-used applications on smartphones (Muscenell & Guadagno, 2012; Oulasvirta, Rattenbury, Ma, & Raita, 2011).

Gender differences also relate to purchase decisions by smart phone users. Wilska (2003) examined the relationship of consumption patterns and mobile phone use when adopting mobile phones as a form of new technology. An “addictive” use of mobile phones linked females to “trendy” and “impulsive” consumption styles. Technology enthusiasm and trend-consciousness related males to consumption and “hard” values.

While currently little research exists that shows a connection between gender and smartphone usage for informal learning and the management of undergraduate coursework, gender differences have been studied in technology affinity. Studies indicate that there are differences of interest in technology and in perceptions of skills toward technology by gender, with women displaying lower levels of interest (Frantom, Green, and Hoffman, 2002). The Technology Affinity Survey (Knezek, 2011) was designed to assess involvement with communications technology, Internet, and media tools. In a study of 68 teacher preparation candidates (Knezek, Mills, Wakefield, & Hopper, 2011) findings showed gender as a significant (p<.05) discriminator for the level of technology affinity. Males were higher in digital technology affinity than females. Morris and Venkatesh (2000) researched the adoption of a new technology in general using the Technology Acceptance Model. Women’s technology usage decisions were strongly influenced by their perceptions of ease of use. In contrast, men’s were more strongly influenced by their perceptions of usefulness.

Gender differences exist in use of social and web based media, consumption patterns, attitudes and affinity toward technology. Results from the Smartphone Usage Instrument (Hopper & Evans, 2012) will explore gender differences in frequency of smartphone tasks and should it prove reliable and valid, could contribute additional findings to the field.

**Method**

In this study, 42 college students enrolled in a blended learning class answered 15 survey questions about how they use their smart phones. Each question asked the students to rate the frequency with which they used their smartphones to complete a specific task for their class by choosing never, once per month, 2-3 times per month, once per week, 3-5 times per week, every day or several times per day. Participants were able to see all of the questions at one time (as shown Figure 1) and experienced no time limits while completing the survey. An assessment of the survey questions was conducted to test for content validity. The survey questions were clear and understandable and passed the fundamental requirement for content validity (Sireci, 1998).

The instrument proved to be reliable, with a Cronbach’s Alpha score of .947 (Cronbach, 1951) (DeVellis, 1991). Dimension reduction through exploratory factor analysis showed two distinct categories of data using varimax criterion (Kaiser, 1958). Two constructs were created based on the categories identified. The first factor centered on using smartphones for logistics and organizational tasks for the course (F1). The second factor centered more on activities that facilitated deeper social relationships, including such tasks as sharing videos or photographs (F2). These categories were also tested for
internal reliability using Cronbach’s Alpha and both factors proved reliable (DeVellis, 1991). F1 returned a Cronbach’s Alpha of .915 and F2 returned a Cronbach’s Alpha of .943. An additional factor analysis which forced 3, then 4 constructs showed no additional factors worth consideration.

Cronbach’s Alpha factor scores demonstrated excellent reliability for creating scales, so scales were created accordingly, defined as F1and F2 (DeVillis, 1991). These scales were then tested for criterion-related validity using a one-way ANOVA test with gender as the discriminating factor. We hypothesized that how frequently students use smartphones to support different informal learning tasks would be dependent on gender. Specifically, we predicted that female students would more frequently use their smartphones in activities related to F2, which helped facilitate social relationships and also involved sharing videos or photographs. The results confirmed our hypothesis.

<table>
<thead>
<tr>
<th>Please indicate how often you use your smart phone for the following tasks in your blended course.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I use my smart phone to check email pertaining to my course from my professor.</td>
</tr>
<tr>
<td>I use my smart phone to email my professor about questions in my course.</td>
</tr>
<tr>
<td>I use my smart phone to Google information I need for my course.</td>
</tr>
<tr>
<td>I use my smart phone to enter due dates on a calendar for my course.</td>
</tr>
<tr>
<td>I use my smart phone to check the course content (Moodle, Blackboard, etc.) for details about my course.</td>
</tr>
<tr>
<td>I use my smart phone to check websites I may need for assignments in my course.</td>
</tr>
<tr>
<td>I use my smart phone to email friends in my class about assignments for my course.</td>
</tr>
<tr>
<td>I use my smart phone to text my professor about assignments for my course.</td>
</tr>
<tr>
<td>I use my smart phone to text my friends about assignments for my course.</td>
</tr>
<tr>
<td>I use my smart phone to call friends about assignments for my course.</td>
</tr>
<tr>
<td>I use my smart phone to read journals or information for my course.</td>
</tr>
<tr>
<td>I use apps on my smart phone to assist in my course.</td>
</tr>
<tr>
<td>I use my camera on my smart phone to complete tasks in my course.</td>
</tr>
<tr>
<td>I use my video camera on my smart phone to complete tasks in my course.</td>
</tr>
<tr>
<td>I use my smart phone to watch videos for my course.</td>
</tr>
</tbody>
</table>

Figure 1: Survey of Smartphone Usage Instrument frequency. ([Hopper & Evans, 2012]

While F1 was far from showing any significant difference based on gender (p=.798), F2 showed a significance score that did merit further consideration (p =.089). Female students used their smartphones more frequently than male students to facilitate social relationships or share videos and pictures. Though this score failed to meet the threshold for significance where p <.05, the sample size was fairly modest. We then tested the effect size to see if a larger sampling should be considered. The effect size is moderately large, with a Cohen’s d =.601 (Cohen, 1988). Due to the moderately large effect size, a slightly larger sample size is worthy of consideration, as this result would be considered educationally significant (Sivin-Kachala, Bialo, & Langford, 1997). By increasing sample size by a modest amount, the
instrument would pass the Criterion validity test, meaning that gender could be a predictive factor in whether the student used their smartphone for high-end social communication and multimedia content.

Conclusion

The Smartphone Usage Instrument tested in this study is reliable. The scales created showed promise in using gender to predict the frequency that undergraduate students will use their smartphones for social relationship tasks and sharing video and pictures in the context of a blended learning class. The effect size is moderately large and with a larger sample size, gender would prove to be a significant predictive factor. Based on these results, we believe this pilot reveals the need for further study on smartphone usage among students in college courses. We intend to make subtle refinements to the instrument, as well as increase our sample size, in an effort to show more conclusively whether gender can be used as a determining factor in predicting the types of tasks students use smartphones to complete when managing college coursework, as well as the frequency with which they complete those tasks.

References


If We Build it, Will They Come?
Designing a Community-Based Online Site for Parents

Sarita Yardi Schoenebeck
University of Michigan
yardi@umich.edu

Amy Bruckman
Georgia Institute of Technology
asb@cc.gatech.edu

Abstract

Parents are busy people. Designing social software for parents requires understanding the particular needs and constraints governing their lives. In this paper, we present a study of a community-based site called ParentNet. Based on prior formative work, ParentNet was designed to support parents in keeping up with their children’s social media use. With 10 months of deployment and 133 participants, ParentNet was successful in some regards and unsuccessful in others. Drawing from log data and focus groups, we arrive at three findings. First, parents may not easily switch from existing school communication platforms that they are already familiar with. Second, school support was critical for promoting adoption. Third, parents felt like they had too much technology in their lives and were not looking for more platforms to keep up with. We conclude with lessons for designing for parents and a discussion of technology overuse as a design constraint.

Keywords: parents, design, social media, youth, schools

Introduction

If we build it, will they come? Designers of social systems ask themselves this question regularly. We operate on the belief that a well-designed, functional system that should be adopted by users. Yet, this is not usually the case. For every successful deployment, hundreds and thousands go by barely noticed. Others experience momentary peaks of interest, then abrupt decline. A small number, like Outlook and Thunderbird, or Facebook and Twitter, or AOL and Skype maintain popularity long enough to leave a lasting impact on individuals and society. These services change the way people interact, communicate, and find information. For some people, especially youth, it’s hard to imagine what life was like before communication technologies. For parents as well, cell phones and laptops have dramatically changed social life in the home.

Yet, there are also growing concerns about multi-tasking, addiction, lack of physical movement, poor relational development, and underdeveloped communication skills in people’s technology use (Carr, 2010; Ophir, Nass, & Wagner, 2009; Sisson et al., 2009). These contribute to a generally shared sentiment that some amount of respite from technology is needed (Woodruff, Augustin, & Foucault, 2007). Parents struggle with questions about whether their family is using too much technology. They also have questions about whether or not to monitor their children and how much to do so (Boyd, Hargittai, Schultz, & Palfrey, 2011; Byrne & Lee, 2011; Rode, 2009). Our formative work observed that parents are challenged to know how to keep up with their children, especially parents who aren’t comfortable with technology (Yardi & Bruckman, 2011).

1Research was conducted while the first author was a graduate student at Georgia Tech.

Acknowledgements: We thank the parents for their time and in particular, the school for their support in conducting this work. The phrase “if you build it, they will come” is a variation of a quote from (Robinson, 1989).
Schoenebeck, S. Y., & Bruckman, A. (2013). If we build it, will they come? Designing a community-based online site for parents. iConference 2013 Proceedings (pp. 54-64). doi:10.9776/13138
Copyright is held by the authors.
In this paper, we describe the design and deployment of ParentNet, an online social network for middle school parents to help them keep up with their children’s social media use. ParentNet was the result of over three years of collaboration with a K-12 private school. It was designed after a fairly extensive amount of fieldwork was conducted with that school which revealed questions and challenges parents faced around managing technology use. ParentNet was intended to give parents a place to keep up with changes in social media and a platform to try out services like chatrooms. Yet, parents engaged with ParentNet and the research team differently than the results of the fieldwork might have predicted.

We present three findings from this research: 1) Parents adopted ParentNet like a mailing list, a communication culture that they were already accustomed to; 2) School support was helpful, and maybe even critical, for adoption of ParentNet; 3) Parents were inundated with technology and weren’t looking for more platforms to have to attend to. Results 1 and 2 extend prior work on deploying systems and motivating adoption but with a new user demographic: parents. Result 3 was a new finding that speaks to a growing movement around technology overuse and refusal. This is the first study we are aware of that focuses on designing a community-based site for parents to manage technology use. We believe the primary contribution of this work provides insights into designing social systems for parents, a very large but potentially unique and understudied demographic of technology users.

Related Work

We focus on two areas of related work. We first review prior work on what motivates people to participate online. This lays the groundwork for understanding challenges in deploying successful online sites. Then, we describe approaches to parenting and consider these approaches as they relate to technology use. We provide this background to explain why we designed ParentNet and also to set the stage for explore how parents engage with a new online site in our research.

Participation in Online Communities

Motivations for participating online have been studied by computer and social scientists for over 20 years. Following Ostrom’s groundbreaking Governing the Commons, a rich new body of research arose that explored challenges in promoting participation and cooperation online (P Kollock & Smith P., 1999; Peter Kollock & Herring, 1996; Wellman & Gulia, 1999). Researchers have drawn on social science theories to understand these motivations and to help designers and developers to understand the successes and failures of systems (Grudin, 1989). Motivations for participating online range from self-interest to altruism (P Kollock & Smith P., 1999). One kind of self-interest is reciprocity, where a person is motivated to contribute in anticipation of future return. Such self-interest can offer benefits to the group which relies on a network of credits and checks. Another example of self-interest is sense of efficacy that helps a person believe she has an impact on the group. Preece and Shneiderman propose a transformation of participation framework called “reader-to-leader” which suggests that successful levels of social participation will involve users moving from reader to contributor to collaborator to leader (Preece & Shneiderman, 2009).

Although quite a lot is known about what motivates people to participate online, translating this knowledge into design decisions has proved to be more challenging. Beenen et al. conducted four experiments to explore whether the uniqueness of one’s contributions might motivate participation and to what extent persuasive messages might do so (Gerard Beenen, Kimberly Ling, Xiaqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, 2004). They found that uniqueness was a motivating factor but most other predictors were disconfirmed. It has also been shown that design decisions might differ based on a user’s attachment to the group versus attachment to the individual (Yuking Ren, Kraut, & Kiesler, 2007). However, invitation styles appear to have little impact on people’s decisions to join a site (Violi, Shneiderman, Hanson, & Rey, 2011). In addition, for those who do choose to join a site, they may continue to participate for different reasons than those that originally led them to the site (Lampe, Wash, Velasquez, & Ozkaya, 2010). In this case, participation between users is usually socially motivated whereas participation to the community is information oriented (Lampe et al., 2010). Designing online communities is too large a body of work to cover here; for a comprehensive study see (Kraut & Resnick, 2012).
To the best of our knowledge, there is no prior work exploring the design of an online site for parents. Though the platform ParentNet is built on is an online social network, the difference is largely a semantic one for our purposes. Our interest is in exploring how to design for a socially and digitally connected set of parents. In the next section, we consider parenting approaches, anxieties, and technology. We do this both to contextualize the goals of ParentNet, and to set the stage for the results and discussion.

**Parenting Approaches**

Attitudes towards parenting began to shift in the early to mid-1900s up to the 1960’s (Nelson, 2010; Pollock, 1983; Stewart & Bond, 2002). Before that change, parents often sought advice from elders in their community, like their own parents or elders at church. They also looked to clergymen for advice on child-rearing. As attitudes and beliefs began to shift, parents began to look to a professional class of experts like doctors and psychologists for parenting advice. With this change came an apparent increase in fear and anxiety among parents, even in cases where there wasn’t strong evidence showing correlation. For example, it has been suggested that fears about not breastfeeding, posture, neighborhoods, exposure to germs, food choices, and a number of other modern features of everyday life related to advice that was coming from this new professional class of experts (Nelson, 2010; Stearns, 2004). Stearns (2004) attributes anxious parenting to a greater sense of vulnerability and frailty in the past century: “Some of our most striking practices, from grade inflation to worries about children's boredom, result from the intersection of beliefs in vulnerability and the influence of wider social institutions” (Stearns, 2004). At the same time that this culture of anxious parenting grew, there was also a growing movement towards children’s rights and encouraging children to develop and express their own feelings (Hefner, 1998). Thus, parents faced an often confusing range of desires to protect their children from disease and sickness while letting them roam freely to explore and learn.

The convergence of the culture of anxious parenting in the past century (Stearns, 2004) with a culture of uncertainty around technology (Borgmann, 1987; Carr, 2010; Lanier, 2010; Nelson, 2010; Turkle, 2011) has led to a complex landscape where parents are both eager advocates of technology in their children’s lives but simultaneously fearful of its effects. The culture of anxious parenting is transferred—and for some parents, magnified—in the technological realm. Nelson (2010) describes a tension in parents’ attitudes towards privacy and surveillance with respect to their children. Specifically, parents, especially middle and upper class parents, suggest they are morally opposed to spying on their children through technological means (drug testing, GPS tracking, spyware), but also acknowledge they would do so if they felt their child was in trouble. Indeed, a complicated menu of options is available for technology use. Consider the following examples from in the context of technology use and whether or not an 11 year old child can join Facebook:

- **Strict parent:** Gets angry and responds that the child cannot join any social networking sites until they say he can.
- **Permissive parent:** Says it is fine by them if the child joins Facebook.
- **Uninvolved parent:** Says they don’t care if the child joins Facebook or not.
- **Authoritative parent:** Asks what the child would like to do on Facebook and agrees to work with the child to use it in appropriate ways.

Authoritative parenting has been accepted by many as the recommended parenting style (Clarke-stewart, 1998; Fagan, 2000; Maccoby, 1992; Steinberg, Darling, & Fletcher, 1995). However, critics have argued that this style was based on studies of White, middle class families and does not take into consideration different cultures, values, and approaches to families (Barber, 1996; Julian, McKeney, & McKelvey, 1994; Rodriguez, Donovick, & Crowley, 2009). Taken together, the parenting landscape is complex, and parents are tasked with the ongoing challenge of setting rules and managing how their children use social media. This challenge has been addressed by a small and growing community of researchers (Boyd et al., 2011; Byrne & Lee, 2011; Lenhart, 2010; Rode, 2009; Yardi & Bruckman, 2011), but a lot of work remains to be done. This section contextualizes our motivations for designing ParentNet, and sheds light on the complexities facing parents raising children in a digital age.
Methods

Participants

We maintained a partnership with a private school in a large urban city in the U.S. for over three years during this project. The school had implemented a one-to-one laptop program (where each student is loaned a laptop from the school for the academic year) among its middle school students (grades 6-8) and was interested in ways of further supporting parents. They were also located near our university and appreciated having a university partnership. Participants were mostly middle to upper class. In general, parents did not suggest that they were early adopters (e.g. waiting in line to buy an iPad) but they were economically able to buy new technologies. Our formative studies showed the ways that parents develop attitudes towards their children’s technology use (Yardi & Bruckman, 2011). In that work, the sample we interviewed was a self-selected convenience sample, chosen as a sample of parents to design ParentNet for.

ParentNet Design

ParentNet was designed to support parents in keeping up with their children’s social media and technology use. It supported these goals by providing up-to-date resources and information about social media and by giving parents a platform to try out services like chat and discuss issues on their mind. We implemented ParentNet building on an out of the box social network platform called Ning\(^2\) with JavaScript, PHP, and mySql extensions (see Figure 1). Some ParentNet features mirror those of well-known social networking services like Facebook. We added custom features like grade-level privacy, school feeds, and back-end analytics for the research team. The left column contained private groups for class of 2015, 2016, and 2017 parents (parents of 6th, 7th, and 8th graders during the 2010-2011 academic year). Below the Groups section was a Members section showing a subset of the network members. Content of the network included updates on new social sites and uses that were coming out (e.g. Google Buzz, Google+, Formspring), updates on technology policy and meetings from the school administrators, and discussions on social media use in the news. A blast notification email was sent to all network members every 1-2 weeks containing a summary of new content.

\[\text{Figure 1. ParentNet screenshot (anonymized).}\]

\(^2\) www.ning.com
During the 2009-2010 school year we conducted a pilot study where 6th grade parents were invited to join ParentNet. This pilot was exploratory and experimental, meaning that it was seen as a trial year before we would be able to gain fuller support of the school. We organized chat help-sessions, provided resources, and observed participation. The goal was to develop a set of design principles around which to redesign the network for the following year. Based on interactions with parents, teachers, and administrators, and observation of use, we redesigned ParentNet for the 2010-2011 year.

We deployed ParentNet again in August 2010. This time the network was seeded with a team of resources, including middle school administration, school IT and technology administrators, and four students from the student “tech team.” We recruited parents by speaking at the school’s opening parent night and one to one laptop information session. The middle school principal also emailed the middle school-wide parent email list inviting parents to join the network. After three months, the parent network had 133 members, 5 of whom were students from the upper school (grades 9-12) “tech team,” 5 of whom were school administrators, and 2 from the research team.

ParentNet Evaluation

To evaluate attitudes towards ParentNet, we conducted four focus groups with parents at the conclusion of the school year (May 2011, one week before the end of the semester). Parents were recruited through emails to the middle and upper school parent mailing lists from the school administration. Two focus groups were conducted with 5-7th grade parents and two with 8-10th grade parents (parents with children in both grade groups were invited to join either group). A total of 28 parents participated. Focus groups were conducted at the school on weekday mornings at different times to align with child drop-offs. The focus groups were structured to complement the network, with a focus on what worked for them and what did not.

We combined ParentNet use with focus group data in the analysis below to show how parents navigate the complexities of social media management. We used a thematic analysis approach to organize focus group transcripts into categories (Boyatzis, 1998). We selected the most commonly recurring themes around ParentNet use and technology attitudes. We then returned to the transcripts and refined the themes.

Results

We begin by describing ParentNet participation and engagement drawing from log data. Then, we draw on focus group data to describe reasons for adoption and non-adoption. Finally, we describe the role of the school in promoting ParentNet and access to participants. In the discussion we reflect on lessons that we can draw from these results.

ParentNet Participation and Engagement

In total, 119 parents joined ParentNet, representing about 30% of the middle school parents at the school. The distribution was skewed towards parents of younger students, particularly parents of 6th graders. This group was new to the middle school, new to the one-to-one laptop program, and likely to be starting to think about buying their children technology like cell phones. Most parents joined the network within the first few weeks of its announcement. The network contained a welcome message at the top for the first month.

A blast notification email was sent to all network members every 1-2 weeks containing a summary of new content. The spikes in Figure 2 show network read patterns as a result of those blasts. In general, parents treated the network like a mailing list with a social presence rather than a social network. That is, they joined, friended other parents, created profile photos, but read content without contributing much of their own. Instead they often shared feedback to the research team in face to face meetings and through the school staff. This indicates that they transferred existing behaviors—email lists from the parent groups at school—to the use of this new platform. We return to this in the discussion. Just under 40% of parents joined ParentNet of the total possible number of parents in the middle school (though for most families, one parent in the family joined which makes the participation rate/family closer to 70%) (see Table 1). As mentioned earlier, ParentNet was advertised by school administration and we believe high join rates were due to school support as well as general interest in the topic of social media. After joining, just over 80% of parents returned to the site at least once and about 30% returned weekly, usually after the weekly blasts were sent out.
Figure 2. ParentNet Traffic for two month period.

Table 1

<table>
<thead>
<tr>
<th>Action</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joined</td>
<td>38%</td>
</tr>
<tr>
<td>Read</td>
<td>81%</td>
</tr>
<tr>
<td>Contributed</td>
<td>29%</td>
</tr>
</tbody>
</table>

Reasons for Adoption and Non-Adoption

During the focus groups, we asked parents about what they liked and didn’t like about ParentNet and what they might want to be different (Tables 2-4). Parents liked its main purpose of helping them keep up with social media that their children were using. They also liked the sense of community that it was designed to promote, bringing together various members of the school community that otherwise would not exist in a single space. One key aspect of this was being able to keep up with school technology policies and updates, which changed frequently. Finally, they generally liked the attention being given to technology in their children’s lives by the school and researchers.

However, parents still did not contribute heavily to ParentNet. When asked what they did not like about it, their responses centered on three interrelated issues: simply not having enough time, having too much communication and coordination in their lives already, and trying to spend less time with technology rather than more. A smaller number said they preferred face to face interaction (talking) or that it was unclear what the benefits were for them.

When asked what could be different or what they wanted, many emphasized that they didn’t necessarily want to discuss what was going on with their children’s technology use. They wanted instructions for how to manage technology use rather than discussing and reflecting on use only (for example, they wanted to know what age their children should be allowed to get on Facebook). However, they also wanted the ability to override any advice that was given to them. What we learned was that they already talked a lot about social media face to face among themselves. They weren’t necessarily looking for more discussion but for firm and tractable advice about what they should do. They also wondered what the short or long-term effects of technology use was and frequently asked us our opinions and for advice (questions that there are unfortunately not many answers for yet).

Some parents wanted the school to help them bear some of the burden of keeping up with technology changes. They wanted the school to make decisions about what children should or should not be doing, but wanted themselves to have full authority to accept or veto such decisions. For example, the school might say students should not be on Google Buzz until a certain age, the way this particular school school openly asserted about Facebook use. This way parents were equipped with a line of arguments such as “school rules say no Buzz until age 14.” Parents thus did not have to explain why the rule is what it is, just that the school says so, (but importantly, parents still had the right to overrule when they want to). Much of the aversion to their children’s use of technology was because parents had a sense that their children—and themselves—were connecting to and through technology too much. For example, most told us they thought their teenagers texted too much, but they did not keep track of how many texts were sent each month from the phone bill. Parents told us that they felt they did not have the energy and resources to figure out what was going on. Some were concerned about what they would have to give up in order to spend more time understanding technology. This category of participants felt that parents who spent a lot of time on Facebook
were making the wrong choice—sacrificing other priorities in order to be spending time online. Thus, we observed that ParentNet was successful in giving credence to the issues parents faced, but it was not successful at giving them readily accessible solutions to these issues. ParentNet was useful for gathering multiple stakeholders online in one place who otherwise wouldn’t have been, such as the school administrators, parents, teachers, and researchers. We report on the role of the school in the next section.

Table 2
What Participants Liked on ParentNet (N=28).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Keep up with changes in social media</td>
<td>24</td>
</tr>
<tr>
<td>Keep up with school technology policy and updates</td>
<td>22</td>
</tr>
<tr>
<td>Community Engagement</td>
<td>18</td>
</tr>
<tr>
<td>Attention being given to the role of technology</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 3
What Participants Did Not Like on ParentNet (N=28).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Trying to spend less time with technology, not more</td>
<td>25</td>
</tr>
<tr>
<td>Not enough time</td>
<td>23</td>
</tr>
<tr>
<td>Too many platforms, hard to keep up</td>
<td>16</td>
</tr>
<tr>
<td>Prefer to talk face-to-face</td>
<td>8</td>
</tr>
<tr>
<td>Unclear what benefits are</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4
What Participants Want (N=28).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Someone else to establish rules about technology use</td>
<td>20</td>
</tr>
<tr>
<td>Effects of technology use (and overuse) to be more clear</td>
<td>17</td>
</tr>
<tr>
<td>Easier way to access information on ParentNet</td>
<td>13</td>
</tr>
</tbody>
</table>

Deploying ParentNet in a School Setting

We present two results here that we think will be valuable to future research. First, working with a school broadened access to participants in ways that a non-school technology deployment would likely not have. Second, ParentNet became a technology probe through which to learn about technology use among schools and families more closely.

Access to Participants

Though these stakeholders shared the goal of educating children, many took ideologically varied approaches to doing so in terms of pedagogy and values. For example, some wanted their children to make mistakes and learn from them—“teachable moments.” Others did not want their children to make such mistakes nor be exposed to factors that could lead to them. Working closely with a school lent us access to this ecosystem and to the many participants involved in it.

In our research the school had numerous stakeholders, including parents, teachers, administrators, technology coordinators, and of course, students. The school provided a central location around which to meet with participants and access to a broad range of parents. When we entered this research we planned to design ParentNet for high school students. After early formative research, we realized that by then technology had already permeated teenager’s lives. Our early interviews with students and parents revealed to us that that the kinds of oversight parents wanted were dramatically different for 10 year old children versus 16 year old children (Yardi & Bruckman, 2011). We found that middle school was the main period of concern for parents. Parents whose children were not yet in middle school (5th grade and younger) said their children simply were not using technology all that much yet. Parents of 10th-12th graders generally conceded independence and autonomy to their children (and we observed that they were more focused on college applications, drinking, and dating at that point). Our formative work with the school—and the school’s own needs—led us to focus on parents of middle school aged children, for whom technology access and use was new and rampant. Easy access to participants enabled us to iterate on our own design and research questions.
Technology Probe

Running a deployment meant there was little control over external variables and we observed many such deviations in our field work. We went into this research taking an action research-like approach, wanting to understand what parents were dealing with and to help make their lives a bit easier with respect to technology in their family's lives. During the year that ParentNet was deployed, there were serious issues relating to sexting and online predators at the school (we learned that these issues seem to be happening at many schools but are rarely publicized in order to protect the school and individuals). As such, parents' attitudes about Internet use varied dramatically depending on how close or far their child might be to these kinds of incidents and simply depending on when during the year we talked to them. As we continued through the complexities of working with parents, understanding school agendas and policies, and the scope of technology use in youths' lives, our intervention became a "technology probe" (Hutchinson et al., 2003) as much as an agent for change. By technology probe, we mean that it was a tool through which to gain more insight into the community and its workings than we would have otherwise. Specifically, it became an artifact through which to understand the differences between what parents said and what they did—an ongoing challenge in designing social systems. We discuss these results in the next section.

Discussion

We find that designing social platforms for parents carries some intriguing challenges. Parents joined ParentNet enthusiastically, but did not participate heavily or continuously over time. We presented three reasons for this. First, school support was useful in gaining adoption for new sites. Second, parents had become familiar with existing tools—especially ones that they associate with the school context—and did not easily switch to new tools. Finally, and perhaps most interestingly, parents were busy and felt they had too much technology in their lives already, what we're calling "technology overuse." We focus our discussion on lessons for deploying parent sites and reflections on technology overuse and parenting.

Lessons for Deploying Community-Based Parent Sites

Use of ParentNet mirrored familiar patterns of participation online where a few parents contributed the most, many read but didn't post frequently, and some joined and did not return (Guzdial & Turns, 2000; Lampe & Roth, 2012; Lampe et al., 2010; Nonnecke & Preece, 2000). The “1% rule,” also known as the 90-9-1 principle, asserts that 1% of people will create content, 9% will contribute, and 90% will lurk (though this omits people who join and rarely or never return). This pattern is seen on popular sites like Twitter where somewhere around 60% of users are estimated to join and then not return; MySpace and Facebook have somewhere around 40% who don't return (Cashmore, 2009). Coordinating multiple stakeholders who are distributed across different locations can result in barriers in use (Lampe & Roth, 2012). Schools, in particular, are complex social institutions. One of the lessons we learned is that we should have done more to change the culture among parents and to better set their expectations for how ParentNet might be used. These kinds of design reflections might be coupled together under the label of “misinformed implementation.” That is, the design specifications we derived from formative work did not easily translate to the best possible design. Beenen et al. speculated this outcome in their own system implementation, suggesting that poor wording, new contexts, and feelings of psychological reactance might have all deterred users from engaging with their system (Gerard Beenen, Kimberly Ling, Xiaoting Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, 2004). It is also possible that parents prefer different software platforms, based on the quality and usability of the system. We are currently designing new technical approaches to supporting parents and children and plan to make the following changes, and offer them here for other researchers looking to do similar kinds of deployments.

1) Design systems for mobile devices. Most parents have cell phones in the U.S. and many have smartphones. Parents are often on the go, picking up children after school and after activities, and we observed those were times that checked their mobile devices frequently.
2) Build plug-ins to existing platforms that parents use. Reduce the number of logins and services that need to be remembered.
3) Leverage an existing community of parents, like a school or other organization. Support from administrators and key leaders help gain buy-in and adoption.
4) Don’t expect to change local culture with just different or newer technology. Parents were accustomed to face to face conversations and school mailing lists and continued to rely on these communication styles.
These are some of the design takeaways that emerged from our research. Many of these speak to broader social issues that we see pervading technology use among parents and children. In the next section we reflect on these issues as they might relate to future design and research opportunities.

**Technology Overuse and the Evolving Landscape of Parenting**

Prior work has shed light on the kinds of worries parents harbor—that their children are circumventing rules about technology use, that they should be monitoring their children’s Internet use more than they do, and that they need to be keeping up with new social media trends (Boyd et al., 2011; Byrne & Lee, 2011; Rode, 2009). But why do these worries surface, especially among families who haven’t experienced anything to worry about beyond normal child development (making friends, school, etc.)? Latour proposed the idea of *instant revisionism* to explain how we develop theories about the ways of the world (Latour, 2004). Problematically, he says, as soon as something surprising or dangerous happens, we look for a quick explanation. We found that some parents strongly disliked FormSpring, Google Buzz, Skype, and other social software for exactly this reason—if they heard about problematic uses of these sites, they immediately worried about their own children’s uses. Thus, decades of theories about children, relationships, and parenting are quickly revised into instant theories about a particular service or platform and its negative effects. Mainstream media is also sensationalist, serving harbingers like the “dark side of Facebook” (Pearse, 2012). Monitoring children is a contentious topic, and suggests helicopter parenting, overparenting, and other kinds of anxious parenting (Stearns, 2004). This seems problematic. Parents need a language and framework to think about their children’s technology use that doesn’t immediately position them as either negligent or hyperparent (see parenting styles in (Baumrind, 1966)). Instead, new theoretical frameworks are needed to talk about how parents can keep up with their children’s technology use while supporting children’s growing autonomy and independence.

The idea of technology overuse and the attention economy is not surprising in today’s culture. Technology ownership is pervasive in the U.S. and in many developed countries. We believe that technology overuse will increasingly have to be a constraint in the design process. Little is understood about how to design for overuse, but two strands from other research might be useful here. The first is the need to design less. That is, to discover when “not to design” (Baumer, 2011). In this approach, researchers are encouraged to reimagine technology not as a solution to a problem, but as complex reconfiguration of a particular setting. Sometimes, designing less or nothing at all might be the best design (Densmore, 2012; Wyche, Tech, & Grinter, 2012). The second factor is what to do when there is “lagging adoption” (Satchell & Dourish, 2009) or when design “goes bad” (Gaver, Bowers, Kerridge, Boucher, & Jarvis, 2009). That is, when a design fails, did designers fail to determine appropriate design guidelines or was a design failure simply unpredictable? What happens next in that situation? Both of these conditions surface a broader set of design constraints than those we discussed in the related work section. For parents—and perhaps for many demographics right now—technology overuse has become an important constraint in the design process. Future research agendas should consider both the social implications of technology overuse, as well as design implications for supporting parents in managing technology use in their lives.

**Limitations**

Schools are rich and complex ecosystems. We chose not to write about or expose the details of major conflicts we observed related to the school or technology use. Our results are focused on our deployment and evaluation but it is important to remember that this is only one piece of a complex puzzle. This work is also the result of a case study—we deployed ParentNet with a particular demographic of parents at a particular school. We anticipate that strong school support enabled the research we did and that working in other schools might deem more challenging. This is the first study we know of in an area that we hope generates more interest and research.

**Conclusion and Future Work**

In this paper we have surfaced areas where designing for parents becomes challenging or bears unique characteristics to social software design. We find that working with schools can help with recruiting and adoption but also requires understanding an existing culture and way of doing things. Most technology research in schools has been related to pedagogy or to student-oriented efforts like encouraging physical activity (Poole et al., 2011). With the increasing pervasiveness of technology across settings—home, school,
and work—we think there is much work to be done to better understand how we might design new sites across these settings. Designing technology for parents also needs to address the growing tension in technology use and overuse among families. Our results also lay the groundwork for future work focusing on social media use among parents. We think designing for parents is an area of important future work and hope this work enables a new genre of parent-focused design research.

References


Jobs and Family Relations: Use of Computers and Mobile Phones Among Hispanic Day Laborers in Seattle

Luis Fernando Baron  
University of Washington  
lfbaron@uw.edu

Moriah Neils  
University of Washington  
nelism@uw.edu

Ricardo Gomez  
University of Washington  
rgomez@uw.edu

Abstract

This paper presents the results of a research project on the use of computers and mobile phones by Hispanic day laborers at Casa Latina, a non-profit organization in Seattle, Washington. Drawing from over 100 interviews, participatory observations and a focus group, we found that information and communication technologies (ICT) help immigrant day laborers to remain connected with their families and their employers. Mobile phones complement but do not replace the use of computers: mobile phones are used primarily to obtain jobs, while computers and the Internet are used mostly to communicate with family and friends. The results of this study offer new insight regarding the way day laborers use ICT to facilitate their navigation and integration into society as immigrants with precarious existences in the US. This study can also inform programs to help provide better support services and training that can effectively meet the needs of these extremely underserved populations.

Keywords: immigrants, day labor, information literacy

Introduction

The US is an industrialized country with high penetration and use of information and communication technologies (ICT). For example, broadband internet reached two thirds of the adult population in the country in 2010, according to Pew Research Center, and cell phones were owned by 88% of American adults in 2012 (Smith, 2010; Smith, 2012). The city of Seattle can be easily regarded as a high-tech city, famous for the presence of Microsoft, Amazon, Boeing and Nintendo among other technology firms. For many in Seattle, life is technology-saturated, and navigating popular culture and public governance requires familiarity with technology tools such as computers and the internet. Not everyone in Seattle is employed in high-tech fields, however, and many of these technological outliers are poor. In 2009, 14% of the population in Seattle lived on an income below the poverty line, and in 2007, 23% of Seattle households reported no Internet usage at home (“Poverty Rate Data,” 2012; CCG Consulting, 2007).

The majority of the poor in Seattle are either American Indian or Alaska Native (29%), Black or African American (21%), or Hispanic or Latino (13%) (“Poverty Rate Data,” 2012). Among the poor in Seattle is a large population of day laborers, most of whom are of Hispanic origin and living in precarious conditions. Many have undependable jobs, low pay, no benefits or health insurance, and live in a state of uncertain immigration status with the looming risk of deportation. According to Valenzuela, Theodore, Meléndez, and González (2006), there are an estimated 117,000 day laborers seeking work on any given day in the US, with an estimated 3,000 located in Seattle. Day laborers most often work in gardening and construction, or as painters, roofers and housecleaners. Their employers are usually either individual homeowners or construction contractors. Immigrant day laborers generally have little formal education and limited English language skills. Employment, when it is available, is often low in pay and of short or uncertain duration; Valenzuela et al. (2006) noted that day laborers rarely make even 15,000 dollars a year, keeping them well below the federal poverty threshold.

Acknowledgements: we wish to acknowledge the workers, staff and volunteers at Casa Latina as well as the graduate students of the University of Washington Information School who participated in this study. Insights from peers and colleagues helped give shape to this manuscript, and blind reviewers sharpened its focus; many meetings at Café Allegro helped to identify and elicit the key ideas we present.


Copyright is held by the authors.
Day laborers are a marginalized population vulnerable to abuse by employers (Valenzuela et al., 2006). Day laborers have found some support in the form of day laborer advocacy centers, now existing in 15 states in the US. One of them is Casa Latina (2012), founded in Seattle, WA in 1994 “to empower Latino immigrants through educational and economic opportunities;” last year, Casa Latina had over 350 active members, dispatched more than 4600 jobs and helped more than 60 of its members to obtain permanent employment. Casa Latina is an active member of the National Day Laborer Organizing Network (NDLON) and other organizations that work to improve the living conditions of immigrant day laborers in the US.

This paper explores the experiences of technology use among some of the poorest and most under-served sectors of American society. In particular, we analyze the use of mobile phones and computers among Hispanic day laborers in Seattle. We seek to offer a more in-depth understanding of the way these extremely marginalized sectors of the population use information technologies to meet their everyday needs. The immigrants we interviewed tend to consider their present situation in Seattle as a transitional period, and many intend to return to their home country after achieving some kind of success in the USA. Day laborers place great importance on mobile phones and computers as tools. They view these items as tools that enable not only survival and adjustment to the challenging situations and environments they are facing in this city and country, but also as devices that will allow them to succeed in realizing the ambitions that enabled their journeys to leave their countries of origin in search of a better future.

The remainder of this paper is organized as follows: we first discuss some of the relevant literature regarding immigrants and day laborers, and the use of ICT among marginalized and underserved communities. We then briefly present the research methods employed in this study, followed by a detailed description and discussion of the findings and their implications. The findings include an analysis of the demographic trends among day laborers and their use of ICT, as well as brief narrative descriptions of the experiences of individual day laborers. Their experiences are then analyzed in regard to their use of computers and mobile phones. Finally, we conclude with pointers for future research in this area of inquiry.

**Literature Review**

This study operates in an interdisciplinary field of information science frequently associated with the “digital divide,” “digital inclusion” or with Information and Communication Technologies for Development “ICTD or ICT4D,” applied to the context of poor immigrants in the US. In a recent examination tracing the evolution of the literature on digital divide, Nemer (2012) remarks that bridging the digital divide will mean much more than just providing neutral access to technology. He states that digital inclusion concerns the concept of an “equal access to public space” as an “essential condition for citizenship.” Access to technology alone is not enough: effective use of ICT in ways that help people meet their information needs and improve their quality of life is essential (Gurstein, 2008).

Ono and Zavodny (2008) studied the extent and causes of inequalities in information technology ownership and use between natives and immigrants in the United States. They showed that assessing digital inequality between immigrants and native inhabitants is important not only because of the significant influence of IT access and skills on employment, educational opportunities and civic engagement, but also because many immigrants - particularly those from Latin America - are disadvantaged relative to those born in the United States. Additionally, Ono and Zavodny (2008) pointed out that ICT skills are crucial to success in the workplace and at school, and that such technical skills play a vital role in civic and political engagement. Furthermore, using data from the Computer and Internet Use Supplement to the October 2003 Current Population Survey (CPS), Fairlie (2007) found that large disparities in home computer and Internet access exist across major racial groups, and that Spanish-speaking Latinos have strikingly low rates of computer ownership and home Internet access. In a study on the gender digital divide in the use of mobile phones in a small community of Latino immigrant farm workers in Southeast Ohio, Garcia (2011) suggests that mobile phones are not inherently empowering to women, and under specific circumstances such as undocumented migration, they can serve as a device that strengthens hierarchical power relations between women and men. However, more in line with our findings, Landry and Kuglitsch (2009) found that a community center that provided ICT access and education to women in Yakima, WA was immensely empowering for the women and their families and
sometimes just “breaking the fear” of turning on a computer is a huge step that should not be underestimated.

The concept of the “embeddedness” of ICT in the daily life of transnational migrants is a rich area of study encompassing studies in communication, development, linguistics, information behavior and other fields. One aspect of this embeddedness focuses on ICT use by immigrants to retain contact and connection with their country of origin. Vertovec (2004) observed that for migrants “transnational connectivity through cheap telephone calls is at the heart of their lives.” He examined the use of both mobile phones and international calling cards and the impact they have on immigrants and their families. Even though a telephone call can’t do everything, it can intensify and ease communication over long distances, and has benefits for those on either end of the line. Leonardi (2003) found that Latinos in the US prefer cell phones to the Internet for interpersonal contact. Participants in his study “did not view computers and the internet as technologies that helped keep people connected.” Benitez studied Salvadoran immigrants in Washington DC and their use of the internet, and found that though the community had limited access, they perceived the internet as a useful tool for family communication (Benitez, 2006).

Some studies explore the use of ICT as a contextualizing tool for societal integration and inclusion amongst migrant populations, either in conjunction with ICT use for contacting their home country, or as a distinct phenomenon. An examination of the literature suggests that the information behavior of immigrants has to take both of these perspectives in hand: “migrants therefore exist in a world of ‘in-betweness,’ negotiating cultural forms and identities at the crossroads of the nation-state and global diasporas” (Srinivasan & Pyati, 2007). Burell & Anderson (2008) particularly explored ICT use among Ghanaians abroad and found their use of these technologies cultivated a connection to the homeland as well as a window into the society they navigated on a daily basis, and to the currents of a wider world. While studying new migrants to New Zealand, researchers found that access to email and the Internet “enabled migrants to make sense of their immigration/adaptation experience, as well as manage daily living” (Holmes & Janson, 2008, p. 51). We discuss elsewhere the emotional barriers faced by Hispanic day laborers to learn basic computer skills (Gomez et.al., 2013).

Operating in this context, we set out to examine how and why the workers at Casa Latina used ICT, their fears and barriers to use, and their motivations and desires. Living within the context of a high-tech city like Seattle, but leading lives that are threatened by day-to-day basic survival needs, what importance does ICT have for this population and how do they use and perceive it? We wish to let the emergent data, and most importantly, the immigrants themselves, speak.

Research Methods

This study explored the uses and perceptions surrounding ICT among day laborers affiliated with Casa Latina, a social service agency in Seattle, WA. Casa Latina had around 400 active members at the time of the study, of which about 75% were men. Between March and June 2012 we conducted structured interviews with 94 day laborers, a focus group with six volunteer leaders, participatory observations of nine introductory computer classes, and six in-depth interviews with trainees in the introductory computer classes, in compliance with University of Washington Human Subjects Division guidelines and procedures. All data was collected in Spanish by native or near-native Spanish speakers.

We conducted in-depth interviews with six day laborers who were attending English classes at Casa Latina. We spoke with three women and three men between the ages of 24 and 62. The interviews focused on their daily routines and the way in which phones and computers were embedded in their lives. We also attended and observed nine computer classes over the course of more than two months.

Representing about 25% of the total members, the 94 structured interviews offer a broad and representative sample of the opinions of day laborers affiliated with Casa Latina. The focus groups, observations and in-depth interviews, on the other hand, offer a deep and rich understanding of the perceptions and experiences of the workers in relation to the topics studied. The combination of both breadth and depth in the data collection enhances the trustworthiness of the findings. An iterative process of analysis inspired by grounded theory was used to elicit emerging themes and trends, which were discussed and probed further among members of the research team, comprised of graduate students under the supervision of a faculty member of the Information School.
Findings: Uses and Users of ICT

According to the survey data, about 75% of the members of Casa Latina are male, and the majority of the workers are between the ages of 41-50. Most originally come from Mexico and Central America (notably Guatemala, Honduras and El Salvador); workers from more than ten countries were included in our survey. Almost half of the workers arrived in the US during the last decade, nearly a quarter arrived during both the 1980s and 90s, and very few came before that. The majority of respondents have attended either elementary or high school and about 10% have attended college. The workers’ job types are highly gender-segregated: men work in gardening, landscaping and construction, and women in house cleaning and, occasionally, child-care.

![Age Distribution](image1.png)

![Level of Education Completed](image2.png)

![Gender Distribution](image3.png)

**Figure 1-3. Characteristics of Day Laborers**

Personal computers (PCs) are used by 67% of the day laborers, though most of them do not own a computer. Most day laborers use shared, public computers at libraries or other public spaces (Casa Latina just recently started to offer limited computer access and basic training). Computers are owned by 39% of the workers or their families, and they are used to connect to the internet either at home (dialup, wireless, broadband) or in public places where Wi-Fi is available (such as a public library or coffee shop/store). Shared computers are available in public places such as libraries, community centers, schools, and even technology stores (some workers noted that they go to the Apple Store or other technology stores to use the computers located there). Finally, in the case of mobile phones, we paid particular attention to the uses of internet-enabled phones (smart phones), which can have data plans with the wireless carrier, or allow their user to connect to public Wi-Fi. Use and ownership of technology tools is summarized in Figure 4.

![ICT Use and Ownership](image4.png)

**Figure 4. ICT Use and Ownership**

Almost half of the participants of the computers classes mentioned that they spend the majority of their time either at home or at a job site, or commuting between the two. The workers have very limited time and resources to get to know, use, and enjoy other places and facilities in the city. Their lives revolve around their homes, jobs and securing food and clothing. Casa Latina is always mentioned as an important reference of their daily life. The survey shows that while 79% of the members go to Casa Latina primarily to seek jobs, only 47% of the members place the highest value of their membership on the jobs they get dispatched: training opportunities are the most important benefit for 27% of the members we
interviewed, while meeting others and building trust are considered most important for 11% and 10% respectively.

Four Experiences of ICT Use

This section introduces four Hispanic day laborers at Casa Latina. In order to maintain authenticity in these narratives, we have decided to retain the uniqueness of actual personal stories, rather than to create fictional characters with aggregated characteristics. The four narratives tend to epitomize salient common patterns in the experiences and uses of communication technologies among the day laborers in our study. The names and other details have been changed to protect their identities.

Betty does not expect to stay in the USA. She uses her basic phone to interact with her employers, using her rudimentary English skills to communicate with them. Because she spends the majority of her time between work and home and has no relatives living nearby, she does not have local social networks of support. Computers offer her a chance to see and be in touch with her family. This technology provides Betty with emotional support, a space for interacting with trustworthy people and the possibility to share immediate daily life with her family.

Adiana feels that she has already accomplished the goals she made when coming to the USA. Even though she would like to stay here, she is considering going back to Mexico to be in charge of her aging mother’s care. Her phone and use of public computers are seen as economic tools, as they help her to find work and maintain a certain level of social life in the United States. Her basic mobile phone allows her to be located by possible employers and a few members of her family. Access to computers also allows her to look for jobs and send in applications via email.

Orlando wants to have a good life here in Seattle, and to be able to maximize his time in the pursuit of this goal. He uses text messages to be sure of the locations and specifics of his jobs. Shared computers give him the latest news on his country of origin. He also thinks that computers can provide him with the information and education necessary to be a better gardener. His occasional phone calls let him not only keep in touch with distant family and friends but also allow him to be in touch with local social networks.

Rafael would not like to stay in the USA. His phone allows him to contact his family, and also to communicate with his grandmother in Guatemala, who provides him with news regarding his old neighborhood. The phone is also his conduit for employment as a day laborer. Learning to use computers gets him one step closer to his dream of designing gardens and supporting a landscaping business. He already has philanthropic plans for the profits from this business; he wants to help fund educational opportunities for young people.

Discussion: The Immigrant Day Laborers’ Experiences of ICT

Our research shows that mobile phones and computers not only provide different methods of societal integration for day laborers in Seattle, but they also affect relationships with their countries of origin. These tools have different meanings and uses; not only do they support physical and material needs (e.g. communication, jobs), they also support emotional and symbolic needs, such as trust, intimacy and empowerment. Moreover, the “embeddedness” of ICT tools within their daily lives is oriented by personal beliefs and aspirations, as well as cultural values. These results give a more complex picture of the immigration processes, and a more nuanced look into the ways that phones and computers are embedded into the daily lives of day laborers.

Mobile Phone Use

Phones represent the most accessible and manageable tool for the majority of the day laborers consulted. According to the survey data, 86% of the day laborers we interviewed and 100% of the women own a cell phone. Only 25% of mobile phone users connect to the internet through their device, though if they do use this feature, most report using it every day. The data of the survey also shows that older members of the group use internet on cell phones for maps, weather, email and info searches/browsing; in addition, younger people use email, Facebook, and YouTube. Younger members of this group, especially men, are more familiar with phones than computers; they consider them easier to operate and useful for social interaction.
“It is useful because the mobile phone puts food on the table” male (44) day laborer from Guatemala
“The phone is the basis of my work. If I don’t load minutes, I feel that I may not do anything.” Female (29) from Mexico

Phones provide them with the possibility to easily interact with employers and to better agree to the terms and conditions of their work. The survey also shows that the strongest benefit of owning a cell phone is the ability to be called by prospective employers. Even though day laborers have to be present at Casa Latina to be dispatched to a job site (there is a raffle in the morning and jobs are assigned from a list of job requests from employers), repeat employment by previous employers is possible and encouraged: having a cell phone where an employer can reach the worker and to make a new contract is indispensable, according to most of our respondents. Even though only 43% of the workers use text messages (SMS) on a daily basis, many of them stated that this tool has provided them with a more agile and effective tool to interact with their employers. The non-synchronous communication allowed by texting gives the workers time to translate their employers’ messages without stress and without “letting on” that they do not understand what is being said. Many interviewees thought that becoming too dependent on a mobile phone was a big risk, but interestingly almost no one considered privacy issues or the possibility of being tracked or located as a risk related to the use of a mobile phone.

Computer Use

“It has being more than 7 years without seeing them… I haven’t seen my children since I left them”. Computer class student.
“Before I was only able to talk on the phone, now I can see my family and it is like I have them very close. It is a very different experience.” Female (43) day laborer from Honduras

According to the survey, computers are less frequently used, and even less frequently owned than mobile phones. While 67% of respondents use computers, only 39% own one, and 46% use a shared computer, mostly at the public library (69% of users of shared computers use them at the library). The qualitative information collected from day laborers shows that the group that owns computers mainly uses them to stay in regular communication with family and friends in their countries of origin—especially through Skype and Facebook. For many of our respondents it is very important to see distant family, to feel closer to them and share daily life experiences, as well as to feel involved in their home country and its issues. Betty, our case study example, is exemplary of the trend.

Many of the workers who attend the computer classes at Casa Latina mentioned that the possibility of seeing and interacting on a regular basis with their distant families has encouraged them to learn to use computers. Visual communication through a service like Skype adds a new dimension to the experience, as they “can perceive other things” through non-verbal cues. In turn their families can also perceive a more complete picture of life in the United States. Visual communication is a personal and intimate activity. It can help explain why the use of shared computers for communication with family and friends is not prevalent and why shared computer users are more concerned with privacy risks, such as people accessing, observing or stealing their information, passwords and identities.

More day laborers use shared computers than own one. The majority of the workers use the computer to look for information regarding international issues and the economic and political situation of their countries of origin, as well as for searching on topics related to their jobs and to personal interests. To a lesser degree, they also expressed that they use shared computers as a source of entertainment, especially to listen to music and to facilitate such hobbies as cooking, sports or even health and beauty.

For this group, the access to computers is related to the idea of progress, and becoming part of a modern society and a more interconnected world. This perception has been reported among users of shared computers in Latin America as well (Gomez, 2012). Frequently during the computer classes, the workers characterized people who know how to use computers as smart, and they noted that those who are able to work and also make money through the use of computers as even more intelligent.

Due to time constrains, many day laborers indicated they are more efficient when using shared computers, especially when they are looking at information about jobs, the second most important use of shared computers in the survey at 33%. During the participatory observations, women in particular
expressed concern about having enough time to go the library to practice using computers. Many women also mentioned that they have had trouble setting aside time to learn computers at all because of their double burden as both mothers and workers.

The participatory observations showed that men seem to have greater access to spaces where they can share tips and information on use of computers, software and applications with their peers. Many of them mentioned that in their social network there are computer experts who have guided and taught them to solve particular issues.

“We can use it to waste time. Before I would go to the park and play basketball and now I’m only on the computer.” Male (50) day laborer from Mexico

The easy access to programs, websites, games with “bad things” and non-constructive information such as porn, prostitution and crime was considered the most negative aspect of computers within the participatory observation group. However, both the groups that owned or shared computers were concerned about the harmful physical effects of computers, notably eyestrain.

Conclusion

While day laborers represent one of the most vulnerable and marginalized populations amongst immigrants in a highly technological city like Seattle, they make extensive use of communication and information technologies. These tools aid them in their integration into the US, as well as help them to maintain ties with social networks near and far. Due to limited economic resources, sociocultural conditions and the difficulties of cross-border travel, mobile phones and notably computers (both owned and shared), provide day laborers the possibility of maintaining a link to the daily life of their far-away families and friends, as well as provide a link to the everyday news and culture of their home towns and countries. Computers also help them be involved in cultural practices and traditions in their home countries (festivals, marriages, celebrations).

Mobile phones are extremely important to day laborers, not only to help them find and keep jobs, but also to negotiate the specifics of these jobs with employers. Especially significant are the uses of mobile phone text messaging, which enable day laborers to have more equal interactions with their employers and empower them as workers. Having a phone allows one to be reachable by employers, and provides access to local social networks, which help to create a strong sense of “connectedness,” confidence and safety. Access to and knowledge of computers facilitates local job searches and provides access to information that is not otherwise easily accessible. This expands the educational possibilities of these workers, and offers new avenues for advancement for the silenced and marginalized groups among immigrant day laborers, in particular women.

While the day laborers we consulted are not universally using the most advanced ICT, they are learning, accessing and employing technology that suits their needs. Their use of ICT allows them to face the labor and personal challenges of their everyday lives. These technologies are helping them to cross the borders of time and place, facilitating a sense of belonging and social integration, and especially helping them to integrate into society through employment. The ICT tools that these workers chose fit well with their provisional and uncertain daily lives. Fulfilling both emotional and existential needs, these technologies are also helping them to make their own lives easier and friendlier, and probably accounts for the perception of ICT as crucial tools for survival and success in a modern city such as Seattle.

Approaching day laborers through the lens of their use of technology has also shed light on our understanding of the daily lives of this particular group of workers, and the complexity of the processes of immigration. Their experiences have helped us to better appreciate their work and the sacrifices they make to ensure a better life for their families in the US or in their countries of origin. Their stories challenge commonly held ideas about the kinds of jobs these immigrants perform, and the strategies they adopt to find and to maintain employment. Moreover, their stories reinforce the idea that immigration is not an individual experience affecting the life of just one person; instead it is a phenomenon that implicates families and whole social networks of local, regional and national communities.

This research has evidenced that mobile phones are not replacing computers as a source of information and entertainment, but that they complement each other in unique ways. Also, we found strong links between ICT use and two related communication practices which are perceived on the same
plane by the day laborers: the development of English language skills and the arrangement of transportation (by bus or by car). Future research can explore these connections further.

A better understanding of the way ICT assists the everyday lives of day laborers can help tailor training programs for immigrant populations in a way that takes into consideration their constraints, their needs, and the emotional barriers they face when making use of technology tools. Moreover many avenues for further research and analysis were identified through this study. For example, the relationships between ICT and the processes of sociocultural integration of immigrants, the role of ICT for the complex and fluid processes of building citizenship as part of the contemporary immigration phenomena, and the importance of ICT for other labor experiences and communities of immigrants, could stimulate new dimensions and studies within the information and immigration fields.

References


Connecting Government, Libraries, and Communities: Information Behavior Theory and Information Intermediaries in the Design of the LibEGov Tool

Paul T. Jaeger
pjaeger@umd.edu

Ursula Gorham
u gorhos@yahoo.com

John Carlo Bertot
jbertot@umd.edu

Natalie Greene Taylor
ngreenetaylor@gmail.com

Elizabeth Larson
elclarson@gmail.com

Ruth Lincoln
rlincoln21@gmail.com

Information Policy & Access Center, College of Information Studies, University of Maryland

Jonathan Lazar
Shutzer Fellow, Radcliffe Institute for Advanced Study at Harvard University and Towson University
JLazar@towson.edu

Brian Wentz
Department of Computer Science and Information Technologies
Frostburg State University
bjwentz@frostburg.edu

Abstract

As e-government grows in scope and complexity, an increasing number of e-government services have surpassed the digital access and literacy of many members of the public. Often referred to as the digitally excluded, these individuals seek information intermediaries such as public libraries and other community anchor institutions to bridge their information needs and e-government systems. In this paper, the authors describe the data surveys, case studies, interviews, site visits, and usability and accessibility testing used to analyze the needs of the public, libraries, and government agencies. The focus then shifts to the methods employed to develop an innovative online resource that supports these information intermediaries and partnerships. Finally, the roles of the online resource in facilitating information access for these information intermediaries are examined through the lens of the theory of information worlds to illuminate the broader implications for this work. These three sections demonstrate scholarship in action - data leading to innovation and wisdom.

Keywords: information behavior, information services

Introduction

Despite the government’s increasing reliance upon e-government to disseminate information to, and otherwise engage with, citizens, many individuals lack the means to access, understand, and use these services. These barriers to direct interactions between members of the public and government agencies have operated to impede the flow of vital information. To overcome these barriers, an increasing number of members of the public have turned to libraries to help them navigate e-government processes.

This paper focuses on the development of one particular resource that seeks to support librarians as information intermediaries and to foster the creation of partnerships to facilitate intermediation and deliver e-government. Research undertaken to date — including surveys, case studies, interviews, site visits, and usability and accessibility testing — is described and analyzed to provide readers with insight into the often divergent needs of members of the public, libraries, and government agencies. Building
upon this research, the goals of this paper are to: 1) review information behavior in the context of e-government; 2) situate e-government within the theory of information worlds framework as a means of understanding e-government information seeking behavior; 3) detail a multi-method approach used in the development of an online resource for librarians to provide e-government services and resources to their users, and 4) draw upon an understanding of e-government information behavior and the theory of information worlds framework to create a model for describing the connections between government, libraries, and communities. Although focused primarily on one e-government resource, there are also broader implications for this work. By examining the development of this resource through the lens of the theory of information worlds, this paper offers a new perspective on how libraries can best facilitate information access between government agencies and members of the public.

**Information Behavior and E-Government**

As e-government has matured into a dynamic socio-technical system encompassing issues of governance, societal trends, technological change, information management, interaction, and human factors (Dawes, 2009), the delivery of information, communication, and services has become one of the central uses of e-government, raising a number of issues in terms of information behavior. Due to the ongoing evolution of this system, an increasing amount of government information and services is now available exclusively online.

From an agency perspective, the benefits are sizeable, particularly in terms of cost-savings and more efficient use of personnel time (Jaeger & Bertot, 2010, 2011). Because of these benefits, many agencies now view e-government as their primary method for interacting with members of the public (Ebbers, Pietersen, & Noordman, 2008; Streib & Navarro, 2006; Bertot & Jaeger, 2006, 2008) and increasingly use social media platforms – such as Facebook, Twitter, and Tumblr – to further these interactions (Bertot, Jaeger, & Hansen, 2012; Bertot, Jaeger, Munson, & Glaisyer, 2010; Jaeger, Bertot, & Shilton, in press).

For the average citizen, e-government access now stands as the primary means of getting government information and interacting with the government (Bertot, Jaeger, Shuler, Simmons, & Grimes, 2009). E-government services are now the vehicle through which citizens accomplish many necessary educational, economic, social, and political functions - immigration and citizenship, social services, voter registration, license application and renewal, tax payments, enrollment of children in school, and many similarly important functions (Bertot, McClure, & Jaeger, 2008; Gibson, Bertot, & McClure, 2009; Holt & Holt, 2010). Not surprisingly, interactions with members of the public are central to both of the primary reasons typically presented in favor of continuing to expand e-government: 1) engaging citizenry in government in a user-centered manner, and 2) developing quality government services and delivery systems that are efficient and effective (Bertot & Jaeger, 2008). The focus, however, has often been on making the interactions easier for the agency, not the citizen (Jaeger & Bertot, 2010).

The challenges that e-government poses for information behavior among members of the public are significant and varied. As of 2012, 20% of adults in the United States are not Internet users, with persons with disabilities, older adults, non-English speakers, those who did not complete high school, and those with low incomes being the most likely to not use the Internet (Zickuhr & Smith, 2012). Even for Internet users, e-government services are often limited by difficulties in searching for and locating the desired information, a general lack of familiarity with the structure of government, lack of education about the value of e-government, language barriers, and attitudes toward technology and government among many users (Jaeger & Thompson, 2003, 2004). All of these significant issues of information behavior have virtually escaped consideration in the development and refinement of e-government in the nearly two decades since the launch of the Web. If they are discussed, they are often presented as a technology-based digital divide, with little attention paid to the inability of certain populations to engage in increasingly digital government services. An awareness of this e-government divide is beginning to build as part of larger digital literacy and digital inclusion initiatives (IMLS, 2011; NTIA, 2011), but widespread awareness remains elusive.

Many members of the public have overcome e-government challenges by seeking information intermediaries to assist them in their usage of e-government. As providers of free public access to the Internet, the public library has served as the primary – or often only available – access point for e-government access, training, and assistance in communities across the country (Bertot, 2009). In a
sense, serving as guarantor of access and training is a natural extension of the established social roles of
the public library. “The public library is one place that is culturally ingrained as a trusted source of free
and open information access and exchange” (Jaeger & Burnett, 2005, p. 487). Given the complexities of
many services, “even if Americans had all the hardware they needed to access every bit of government
information they required, many would still need the help of skilled librarians whose job it is to be familiar
with multiple systems of access to government systems” (Heanue, 2001, p. 124).

Several years ago, it became commonplace for government agencies to direct users with
questions about their websites to the local public library (Bertot et al, 2006a, 2006b; Bertot, McClure, &
Jaeger, 2008). Now, many federal, state, and local government agencies continue to rely upon public
libraries to provide residents with access to and guidance in using e-government, directing residents to
the nearest public library to obtain this assistance (Bertot et al, 2006a, 2006b; Jaeger, 2009; Jaeger &
Bertot, 2009). For example, the Florida Department of Children and Families reduced the number of case
workers and assistance providers by more than 3,000 positions due to its implementation of the
AccessFlorida online application system, resulting in near complete lack of available agency staff from
which users can seek assistance (Gibson, Bertot, & McClure, 2009).

Government agencies indicate that relying on libraries for e-government access and assistance
allows the agencies to focus on other issues (Bertot et al., 2006a, 2006b; Fisher, Becker, & Crandall,
2010). However, the provision of access assistance by public libraries creates a range of service, funding,
technology, and political challenges for the libraries (Bertot & McClure, 2007; Jaeger & Bertot, 2011).
While the E-government Act of 2002 included language regarding the need for federal government
assessment of the impacts of the law on public libraries and other social institutions, such studies were
never funded or conducted. As a result, the advent of e-government has raised enormous issues of
information behavior within the context of government information, communication, and services that
have yet to be adequately addressed.

The Theory of Information Worlds

In designing the goals and method for this project, the authors opted to employ the theory of
information worlds as the conceptual framework for understanding the roles of information behavior in
e-government usage. The theory of information worlds provides a framework for understanding and
studying the multiple interactions between information, information behavior, and the many different social
contexts within which it exists – from the micro (small worlds) to the meso (intermediate) to the macro (the
lifeworld).

Building on the work of Elfreda Chatman and Jurgen Habermas, the theory argues that
information behavior is shaped simultaneously by both immediate influences, such as friends, family, co-
workers, and trusted information sources of the small worlds in which the individual lives, as well as larger
social influences, including public sphere institutions, media, technology, and politics (Burnett & Jaeger,
2008, 2012; Jaeger & Burnett, 2010). These levels, though separate, do not function in isolation, and to
ignore any level in examining information behavior results in an incomplete picture of the social contexts
of the information. As such, scholarly explorations of information behavior should account for the different
levels to fully understand the social drivers of information behavior and the uses of information in society.

The theory of information worlds seeks to expand the perspective brought to studies of
information behavior in society and to increase understanding of the myriad ways in which information
plays a significant role in social, political, and personal lives. In examining these social structures, and the
ways in which they constantly interact with and reshape one another, the theory of information worlds
focuses on five social elements:

- Social norms: a world’s shared sense of the appropriateness of social appearances and
  observable behaviors;
- Social types: the roles that define actors and how they are perceived within a world;
- Information value: a world’s shared sense of a scale of the importance of information;
- Information behavior: the full range of behaviors and activities related to information available to
  members of a world; and
- Boundaries: the places at which information worlds come into contact with each other and across
  which communication and information exchange can – but may or may not – take place.
As with the social structures within information worlds, the elements are interrelated and constantly interact with and influence one another.

As localized information worlds, each small world has its own social norms, social types, information behavior, and understanding of information value. The members of each small world have established ways in which information is accessed, understood, and exchanged within their world and the degree to which it is shared with others outside the small world. A typical person is a part of many small worlds – friends, family, co-workers, people with shared hobbies, and so forth - and there is no real limit to the number of small worlds to which an individual can belong. Only in extreme circumstances of social isolation can an individual exist in only one small world.

Any one of these small worlds may offer many places where its members are able to interact with members of other small worlds. Information moves through the boundaries between worlds via people who are members of two worlds or through interaction between members of two small worlds in a place where members of different small worlds are exposed to other perspectives. Further, the contact between small worlds and other inputs from the lifeworld can lead to the creation of new worlds as information passing over the boundaries between worlds either blurs those boundaries or otherwise transforms or changes information behaviors and perceptions of information value. Encountering other small worlds can occur through public sphere institutions, such as in a public library, or through new technological avenues of communication and exchange, such as social networks on the Internet. As information moves through boundaries between small worlds, the information is treated, understood, and used differently in each small world in line with the social norms of that world. As a result, the same information may have a different role within each small world.

Together, these small worlds constitute the lifeworld of information. The way that, as a group, the small worlds in the lifeworld treat information will shape how the information is treated within the lifeworld as a whole. As the information moves between small worlds, more and more small worlds will decide how to treat this information, generating an overall perception of the information across the lifeworld. The more small worlds that are exposed to information, the more exchange between small worlds there will be, and the better chances there will be for a democratic perception of and approach to the information.

However, beyond the small worlds, there are also influences at play in the lifeworld that shape the way that small worlds perceive information. Some of these influences increase contact between small worlds and promote democratic engagement in the lifeworld. Libraries, schools, and other public sphere organizations exist specifically to ensure that information continues to move between the small worlds and members of each small world are exposed to other small worlds. In sharp contrast, other influences serve to constrain the movement of information between small worlds or constrict the socially acceptable perceptions of information. The most influential information worlds – such as those who possess political power or those who control the media – can use their power to push back against the collective small worlds to enforce a minority perception on the majority, asserting control over the information in the lifeworld.

Some influences on small worlds and the lifeworld are inherently neutral, but can increase or decrease information access and exchange. ICTs, such as the Internet and online social networks, act as a way for small worlds to connect in new ways and to reach other small worlds that would not otherwise touch their boundaries but they can also work to homogenize perspectives or enforce hegemonic perspectives of small but powerful small worlds on the lifeworld. In total, the small worlds are shaped by all of these larger influences, but also have the power collectively to define the parameters of the external influences.

The theory of information worlds, thus, attempts to account for all of these social and structural elements at work in the shaping of information behavior within a society. While there is obviously great benefit in studying the ways in which information behavior is shaped by the micro, the meso, or the macro level, studying them across levels will provide a much richer and more nuanced understanding of the ways in which information is perceived and moves through society. Though the theory of information worlds presents a much more complex approach, it is intended to compensate the researcher by providing a more thorough and realistic picture of the issues being studied.
LibEGov Project Methodology

The study’s methodology was designed to account for the micro, the meso, or the macro level influences on the information behaviors of members of the public, public librarians, and government agency employees in using e-government to deliver e-government. The data collection efforts, which examined information behavior considerations from the perspectives of members of the public, government agencies, and public libraries, sought to explore the following research questions: 1) What e-government service roles do public libraries provide to their communities? 2) What partnerships have libraries formed with government agencies in the provision of e-government services? 3) What are the success factors and/or barriers to forming partnerships with government agencies? 4) What are the challenges that libraries face in serving as e-government providers? To answer these questions, the study used an iterative multi-method design that provided both qualitative and quantitative data regarding the information behavior context of e-government services.

The first step in the data collection was to gather the agencies’ perspectives – what they wanted to accomplish through e-government and how they designed their online presence to account for information behavior of members of the public. Representatives of five federal agencies were interviewed, and the findings from these interviews provided the government perspective on the process of e-government intermediation.

The next data collection method gathered information about the information behavior of members of the public when seeking to interact with e-government through libraries. The 2011-2012 Public Library Funding & Technology Access Survey (PLFTAS) is the 14th in a series of public library Internet access surveys conducted since 1994 (Bertot et al., 2012). The survey was conducted between September 2011 and November 2011, yielding 7,260 responses, a response rate of 83.1%. The survey drew a proportionate-to-size stratified random sample that considered the metropolitan status of the library (i.e., urban, suburban, and rural). More specific methodology issues regarding the survey are available at [http://www.plinternetsurvey.org](http://www.plinternetsurvey.org).

Building upon the PLFTAS data, the third stage of data collection sought to enhance the understanding of e-government-related information behavior of members of the public through site visits to seven public libraries in five states (Connecticut, Florida, Georgia, Maryland, and Texas). The site visits included interviews with state library staff, government officials, and community organization leaders, all of which occurred between April 2011 and October 2011. For this component of the study, the authors sought to include geographically dispersed public libraries that serve a diverse set of communities (e.g., rural, suburban, and urban; high immigrant concentrations; underserved populations; high poverty). Library characteristics, such as size, number of staff, and known e-government partnerships, also factored into the selection process. In advance of the site visits, research was undertaken to identify existing e-government collaborative approaches, leading to a preliminary assessment of “best practices” that guided the development of the online resource.

The aggregate and generalizable data regarding public library e-government service provision and challenges generated from the survey, together with the on-the-ground assessments of library and government collaborative efforts offered by the site visits and interviews, provide rich insight into the context in which libraries are currently providing e-government services. Based upon the findings from these interrelated data collection efforts, the authors were able to identify key elements to be included in an e-government Web resource (LibEGov), initially focused on the areas of immigration and taxation. The LibEGov resource is intended to foster the role of public librarians as an information intermediary between members of the public and government agencies. This is accomplished through provision of references, best practices, resources, social media, and interactive tools to facilitate intermediation and the delivery of e-government services. Central to LibEGov are a number of features intended to support information behavior that leads to successful engagement with e-government, including: 1) detailed guidance on how to conduct a community needs assessment; 2) a tool to help librarians better understand complex immigration processes; 3) strategies for identifying and reaching out to potential partners; and 4) a community forum that encourages users to share their e-government experiences with one another.

Upon conclusion of the data collection efforts, the project team began the development of the Web resource’s site design and content. Iterative usability testing during the development and refinement of the system has served to ensure that the perspective of the public librarians – the intended users of the site – was central to the completed LibEGov resource. The usability testing was conducted at public
libraries, allowing librarians to test sample queries in the resource in the setting in which they will be using the resource to address information needs of members of the public. In line with best practices in usability and accessibility testing (Dumas & Redish, 1999), the task lists and survey instruments were pilot tested before data collection began. For each of the two rounds of usability testing, 35 library staff evaluated the LibEGov web resource, representing libraries in urban, suburban, and rural areas. These numbers exceed the generally accepted number of participants in usability testing (Lazar, Feng, & Hochheiser, 2010; Turner, Lewis, & Nielsen, 2006). The first phase focused on usability tasks in the following four areas: user account registration and functionality, information seeking, bookshelf functionality, and blog/forum usage. The second phase occurred after additional content had been added to the site, so it focused primarily on information seeking behavior.

Additionally, two experts in interface accessibility individually used a structured, 3-phase approach to assessing the accessibility of LibEGov: listening to a web page with a screen reader while visually viewing the page, using typical non-visual navigational techniques, and code inspections to ensure compliance with each paragraph of the Section 508 interface regulations. This widely accepted approach ensures compliance with the law, addresses different types of disabilities (for instance, since the keyboard access required by screen reader users is also what users with motor impairments need), and allows for multiple viewpoints of the accessibility features of the website (Lazar et al, 2010; Lazar, Wentz, Biggers, et al, 2011; Lazar, Wentz, Bogdan, et al, 2011). In anticipation of a wide release of the resource during the summer of 2013, the next phase of the project will include field testing of LibEGov through workshops at which additional feedback regarding content and design will be solicited from public librarians and employees of government agencies.

From Theory to Method to Design

The methods used in the data collection undertaken to inform the design of the LibEGov resource employed a range of concepts from the theory of information worlds. The survey and case studies collected information on the information behavior of members of the public in the library, capturing the micro level influences and small worlds behavior. These same data collection methods, in conjunction with the usability testing, collected the information behavior of the librarians in their roles as intermediaries, recording the intermediate, meso level influences and behavior. The interviews with government agencies focused on the macro level influences and lifeworld behavior that frame the e-government-related behavior at the other levels. The ultimate goal for LibEGov is to serve as a boundary – a place at which communication and information exchange related to e-government can take place across the small worlds of members of the public, public librarians, and government agencies. In order for LibEGov to function in this capacity, however, it must account for the information behaviors of each relevant party. Through a synthesis of the findings from various data collection efforts, a clear picture of the different information behaviors that needed to be captured by LibEGov has emerged.

As the public library serves as the intermediary between members of the public and e-government, the emphasis in each data collection effort was placed firmly on information behavior in the library. The members of the public comprise one key group of small worlds, as the different small worlds in a community may have different information behavior and intermediation needs in relation to e-government. The public librarians, in trying to meet these e-government needs, also stand as their own small world of information. Thus, in the context of e-government, the library is not only ensuring information flows between various small worlds in the community; it is ensuring information flows between community members and the government.

The data collected revealed a great deal about the social norms, social types, information value, and information behavior of members of the public in seeking e-government information, communication, and services. It is entirely within the social norms and social types of the small world of public librarians to try to meet these e-government needs, as evidenced by the vast majority of public libraries that report providing an array of e-government services and support (see Table 1). A wide range of questions in the survey and the case studies revealed how small worlds are seeking help with e-government in libraries, the specific information problems they are seeking to address, and the types of assistance needed to promote effective information behavior. Hartford Public Library (HPL) in Connecticut – one of the libraries at which a site visit was conducted – has developed a comprehensive approach to addressing these issues in the realm of immigration. The American Place (TAP), through a partnership with United States
Citizenship and Immigration Services (USCIS), makes effective use of the library’s in-house resources to help an increasingly diverse group of recent immigrants secure citizenship and develop its English language skills by offering English as a second language (ESL) classes, life skills workshops, instruction on the use of the Internet and various software, and assistance with applying for the annual U.S. State Department visa lottery. Viewed through the lens of the theory of the information worlds, recent immigrants are a small world with very distinctive information needs, as well as unique social norms, social types, information value, and information behavior.

Table 1
E-Government Roles and Services of the Public Library Outlets, by Metropolitan Status

<table>
<thead>
<tr>
<th>E-Government roles and services</th>
<th>Metropolitan Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Urban</td>
</tr>
<tr>
<td>Staff provide assistance to patrons applying for or accessing</td>
<td>97.3%</td>
</tr>
<tr>
<td>e-government services</td>
<td>(n=2,131)</td>
</tr>
<tr>
<td>Staff provide as needed assistance to patrons for understanding</td>
<td>93.6%</td>
</tr>
<tr>
<td>how to access and use e-government Websites</td>
<td>(n=2,050)</td>
</tr>
<tr>
<td>Staff provide assistance to patrons for understanding government</td>
<td></td>
</tr>
<tr>
<td>programs and services</td>
<td>57.8%</td>
</tr>
<tr>
<td></td>
<td>(n=1,266)</td>
</tr>
<tr>
<td>Staff provide assistance to patrons for completing government</td>
<td>71.0%</td>
</tr>
<tr>
<td>forms</td>
<td>(n=1,555)</td>
</tr>
<tr>
<td>The library developed guides, tip sheets, or other tools to</td>
<td>33.6%</td>
</tr>
<tr>
<td>help patrons use e-government Websites and services</td>
<td>(n=735)</td>
</tr>
<tr>
<td>The library offers training classes regarding the use of</td>
<td>24.5%</td>
</tr>
<tr>
<td>government Websites, understanding government programs, and</td>
<td>(n=537)</td>
</tr>
<tr>
<td>completing electronic forms</td>
<td></td>
</tr>
<tr>
<td>The library partnered with government agencies, non-</td>
<td>43.1%</td>
</tr>
<tr>
<td>profit organizations, and others to provide e-government</td>
<td>(n=941)</td>
</tr>
<tr>
<td>services</td>
<td></td>
</tr>
<tr>
<td>The library worked with government agencies (local, state, or</td>
<td>12.3%</td>
</tr>
<tr>
<td>federal) to help agencies improve their Websites and/or</td>
<td>(n=270)</td>
</tr>
<tr>
<td>e-government services</td>
<td></td>
</tr>
<tr>
<td>The library has at least one staff member with significant</td>
<td>31.4%</td>
</tr>
<tr>
<td>knowledge and skills in provision of e-government services</td>
<td>(n=687)</td>
</tr>
<tr>
<td>Will not total 100%, as categories are not mutually exclusive</td>
<td></td>
</tr>
</tbody>
</table>

In an effort to make immigration information more easily accessible to those individuals who most need it, HPL created an electronic information center that provides links to the most frequently requested items on the USCIS website (e.g., forms, online case status check), as well as to legal and support service providers (Naficy, 2009). By identifying these needs, as well as the problems that recent immigrants face in meeting these needs, TAP has been able to serve as an effective intermediary by facilitating unimpeded information flows between this particular small world and the government.

The data collection confirms the long-standing belief that the public library is the one community institution in which marginalized small worlds in the community – to which individuals of low socio-economic status, education and/or literacy levels, older adults, and non-English speakers belong – can seek and receive help interacting with e-government. The emphasis placed on e-government services and support also demonstrates that e-government is given high information value within the small world of public librarians. The current information behavior of public librarians, however, demonstrates a clear
need for a tool such as LibEGov to support their ability to meet the vast and complicated information needs related to e-government.

As reported on the 2011-2012 PLFTAS survey, overall, adequate staffing was the greatest challenge faced by libraries in meeting patron e-Government needs (44.9 percent). E-government expertise was also a challenge, with 50.5 percent of libraries reporting that library staff does not have the necessary expertise to meet patron e-Government needs. This lack of expertise manifests itself in various ways, including the inability of library staff to guide users through e-government transactions governed by often-complex legal and regulatory implementations (e.g., citizenship). The experience of Austin Public Library (APL) in Texas – another site visit library – highlights several of the challenges that public libraries face in providing e-government services. APL has been operating New Immigrant Centers at multiple branches, offering materials for citizenship and English language study in multiple languages and formats since 2000. In recent years, however, the delivery of these services has been hindered by, among other factors, the inability of the library’s resources to effectively meet the needs of immigrants. By way of example, websites and web resources have yet to be translated into appropriate languages, due to constraints on staff time. Here, the small world of librarians is aware of the information needs of the small world of recent immigrants, but there are information barriers in place that prevent the necessary information from passing over the boundaries between these two worlds.

Using the theory of information worlds to better understand the information behaviors of members of the public as they seek to interact with e-government, as well as of public librarians in their role as e-government intermediaries, LibEGov has been conceptualized as a boundary over which e-government information, communications, and services pass to each of the small worlds touched by e-government. Content in the areas of immigration and taxation were targeted to librarians and the small worlds they represent. Designed for users to pick and choose relevant content, the site recognizes different user groups and encourages librarians to learn more about their community through needs assessments and the discovery of relevant partners in their areas. At the same time, the site attempts to present this range of content as a continuum of services, ranging from ready reference materials librarians traditionally use for quickly looking up facts and figures to suggestions on how to implement subject-specific outreach programming in the users’ own libraries. The use of social media by government agencies has been highlighted as a way to thin the boundaries between the agencies and librarians, acting as both news delivery and feedback mechanisms.

The bookshelf feature was designed for users to gather relevant content for future reference. A forum designed for interaction among librarians and between librarians and agency representatives encourages meso-to-meso discussions, as well as meso-to-micro interactions. Posting to the forum requires logging in, but anyone may read through the threads. Forum users are encouraged to submit practices that have worked in their library or common issues they have come across in the course of delivering e-government services, allowing small worlds to fit into the lifeworld of e-government.

Conclusion

The primary goal of the research into LibEGov has been the development of a tool to facilitate information access between government agencies, information intermediaries, and members of the public. However, as this paper details, the lens of the theory of information worlds can also provide a new perspective on the facilitation of information behavior that leads to successful interactions with e-government information, communication, and services by diverse populations within a community.

Through the information worlds theoretical lens, the study offers a range of findings that inform the development of the LibEGov resource. These include:

- In order to serve as a useful and effective resource, LibEGov needs to bridge the multiple small worlds of agencies, librarians, communities, and users. Though the site is not intended to serve the public directly, it must help librarians navigate the small worlds of multiple user communities that require a range of assistance in order to engage in e-government.
- Librarians, agencies, and users each exhibit different information behaviors, and assume different roles vis-à-vis these behaviors, which can create conflict among these groups. For example, agencies are often concerned with the efficiency of their e-government services, particularly costs. This often results in a transactional approach that emphasizes legalistic and perfunctory information dissemination. Users, on the other hand, are concerned with effectiveness and ease
of use so as to meet their information and e-government needs. Serving as intermediaries, librarians are concerned with both efficiency and effectiveness, and are often forced to fill gaps created by the implementation of e-government services by agencies and the inability of users to successfully engage with these services.

- Intermediaries may not interact with one another as they focus upon meeting the information needs of the particular small worlds they serve, leading to isolation in terms of their e-government service operations. The breadth of e-government information that intermediaries must track, however, makes meso-to-meso interactions vital to the control of this particular service area. Putting the small worlds into the context of the large e-government lifeworld encourages collaboration and creating a forum in which this interaction can take place is one step in the progression of increased interaction among intermediaries.

- The library as an intermediary institution has a unique opportunity to encourage small-world interaction and to change the way the lifeworld of government information is disseminated. While users are segmented into their own groups, libraries offer a place to bring users with similar needs together, potentially creating a critical mass that can influence the agencies providing the information. Librarians in this respect can act as change agents in the lifeworld and as representatives of individuals from various small worlds, but only if agencies give them the opportunity to voice their observations.

- The library also has the opportunity to create connections between and links within small worlds that can foster interactions, communication, and lessen isolation. This can create broader communities to facilitate user, agency, and library e-government efforts.

- Libraries should not assume that others are aware of the place they occupy within and among information worlds. They must continue to promote the various ways they are already facilitating flows of information between the small worlds they serve and government agencies so that resources such as LibEGov reach the greatest possible audience.

- Agencies must recognize the existence of the small worlds of user groups and ensure that information is available in different formats, languages, and venues. Along with varying information behavior, small worlds will have different cultural expectations, different attitudes about e-government, and different means of accessing the information. By connecting with libraries through LibEGov, agencies can learn more about the diverse e-government needs of various small worlds, with the goal of providing information through different accessible means and supporting libraries in their efforts to transfer information to their users via these different means.

These same lessons can also help in broader considerations of connecting government, intermediaries, and members of the public through e-government in future research and subsequent development of online tools. For the vast majority of e-government projects, the focus has been on a "one size fits all" approach to user groups, with little thought to differing needs of user populations (Bertot & Jaeger, 2006, 2008). Yet, as this study shows, information behavior theory can help to connect e-government and its stakeholder groups. Beyond the theory of information worlds, there are many other theories of information behavior that can be applied in the context of e-government access and usage (Spink & Heinstrom, 2012). Ultimately, creating linkages between theoretical perspectives on information behavior to e-government information, communication, and services may prove to be instrumental in efforts to overcome many of the barriers to e-government usage currently faced by different populations.

References


Abstract

This paper offers a framework for analyzing and comparing privacy and privacy protections across (inter alia) time, place, and polity and for examining factors that affect privacy and privacy protection. This framework provides a way to describe precisely aspects of privacy and context and a flexible vocabulary and notation for such descriptions and comparisons. Moreover, it links philosophical and conceptual work on privacy to social science and policy work and accommodates different conceptions of the nature and value of privacy. The paper begins with an outline of the framework. It then refines the view by describing a hypothetical application. The paper concludes with an argument that the framework offers important advantages to privacy scholarship and for privacy policy makers.

Keywords: privacy, information privacy, comparative privacy, privacy policy

Introduction

There is a substantial popular concern about privacy in light of technological advances, greater sharing of information via social networks, and increased power of state and non-state actors to collect information about individuals and institutions. That concern coincides with a growing body of privacy scholarship spanning a broad range of disciplines. One area of inquiry concerns making comparisons of privacy and protections in different places or at different times, for example across national boundaries (Regan 2010; Bennett 1992; Altman 1977; Spiro 1971). A related line of inquiry concerns contextual factors that affect privacy protections and privacy rights. For example, whether the search of one’s briefcase constitutes a privacy violation depends on the setting in which it occurs: an airport security zone, a public sidewalk, or elsewhere (Nissenbaum 2010). Although these lines of scholarship are important and growing, they are in their early stages. This paper seeks to advance the scholarship in making privacy comparisons by providing a framework and conceptual foundation for defining and identifying aspects of privacy and its context in order to better analyze privacy, privacy protections, and privacy rights.

The framework we provide accomplishes several things. First, and most important, it provides a way to describe with precision particular aspects of privacy and privacy’s context. It also allows one to compare privacy in different settings according to variables such as time, location, and polity. To do so, it provides a flexible vocabulary and notation to facilitate such descriptions and comparisons. The framework provides, so far as possible, a morally neutral way of describing and comparing privacy states, and hence does not assume the answers to any questions about the moral importance of privacy in particular cases. Finally, the framework provides a way to link philosophical and conceptual work on privacy to social science and policy work by providing a tool for describing and comparing privacy that both instantiates aspects of the philosophical literature and can accommodate different conceptions of the nature and value of privacy itself.
The paper begins with a discussion of a number of conceptions of privacy and privacy protections. It will argue that those conceptions are incomplete, fail to capture the full range of possibilities for the state of personal privacy, and do not allow for comparisons of privacy states. We offer instead a conception that focuses on privacy as a three-part relation between some individual or institution, some domain of information, and some other individual or institution with respect to whom the first has (or lacks) privacy. Put another way, the three-part relation is a general feature of privacy, and any privacy state (i.e., state of affairs regarding the privacy of some individual or entity) can be expressed in terms of that three-part relation. Rather than arguing for a particular conception of that relation, our view is compatible with a broad range of views about the nature and value of privacy. After setting forth this conception, we further specify the view by describing a hypothetical comparison across states.

Privacy and Context

Several things motivate this paper. One is scholarly interest in comparing the laws and norms protecting privacy across different states or countries. Relatedly, there are important questions surrounding the explanations for such differences. To understand how and why countries offer different types of privacy protections it will be useful to have a framework to systematically describe those differences. This interest in differences in privacy norms is manifest in recent scholarship focusing on privacy’s context. In her recent book Privacy in Context, Helen Nissenbaum makes the case that privacy norms must be understood in terms of “contextual integrity.” She argues that privacy losses are distressing when they violate informational norms, which is to say when they violate norms restricting flows of information. Important here is that those norms are “systematically related to characteristics of the background social situation.” (Nissenbaum 2010, 129) She maintains that “disparities across societies, cultures, and historical periods may manifest in differences” in privacy and informational norms. (Nissenbaum 2010, 134–35) Despite this emphasis on the social factors affecting informational norms, Nissenbaum leaves to “empirical social scientists” the question of how societal and cultural factors give rise to privacy and informational norms.

This paper takes up the task of understanding privacy’s context in three ways. First, it specifies several relevant aspects of any privacy context. Second, by imposing a structure on analyzing privacy it allows for comparisons across “societies, cultures, and historical periods” that may have different privacy norms, and allows one to analyze underlying causes of such differences. Thus, third, the framework provides a tool to predict what privacy norms will be and how they will change.

The View

The foundation of our framework is that any conception of privacy must account for three things that stand in some relation to one another. So, for example, Martijn Blaauw argues that privacy is fundamentally about some person or persons, some set of propositions about the first person, and some other person or persons who know, or do not know, the propositions in the set (Blaauw n.d.). On this view, in order to understand Zeke’s privacy in health information, we must account not just for Zeke, but also for some set of propositions regarding Zeke’s health (e.g., propositions regarding Zeke’s medical history, physiological traits, habits, and so forth) and for some other person or persons who knows, or does not know each of the propositions regarding Zeke’s health. The key point here is that simply describing Zeke as having or lacking privacy is incomplete without specifying the range of propositions regarding which he has (or lacks) privacy and the other persons with respect to whom he has (or lacks) privacy. This is important, for one will often have privacy in some respects but not others. Zeke may have privacy regarding the set of propositions concerning his health with respect to his coworker, but not with respect to his insurer. And he may lack privacy regarding the set of propositions concerning his health with respect to his insurer but retain privacy regarding the set of propositions concerning his reading habits with respect to his insurer. Understanding privacy as a three-part relation forces us to be specific.

A related account is proffered in (Rubel 2011). Like Blaauw, Rubel argues that privacy should be understood as a three-part relation, though he articulates the relevant parts differently. On this view any particular instance of privacy must involve some person or persons P, some domain of information O, and some other person or persons Q. And for P to have privacy regarding O with respect to Q is for Q’s ability to make reasonable particularized judgments about P regarding O to be limited (Rubel 2011, 278–79).
Important for our purposes here is that by understanding privacy as necessarily involving three parts, we can use an expression such as \( POQ \) to denote any privacy instance or privacy state.

The difference between the Blaauw account and the Rubel account concerns the nature of the privacy relation. On Blaauw’s view, privacy is a knowledge relation. If we let \( O \) be the relevant set of propositions concerning \( P \), on Blaauw’s view, \( P \) will have privacy regarding \( O \) with respect to \( Q \) if, and only if, \( Q \) does not know the propositions in \( O \). In contrast, on the Rubel account privacy is about reasonable, particularized inferences, such that \( P \) has privacy regarding \( O \) with respect to \( Q \) to the extent that \( Q \)’s ability to make reasonable particularized judgments about \( P \) and \( O \) is limited. Suppose, for example, that \( Q \) reads \( P \)’s medical record, which states that \( P \) has Lyme disease. It would under normal circumstances be reasonable for \( Q \) to make the inference that \( P \) has Lyme disease, and hence \( P \)’s privacy regarding her health information \( (O) \) decreases with respect to \( Q \). However, because \( Q \) can make such judgments without actually knowing propositions within the domain \( O \), the Rubel account will recognize some cases as privacy losses that Blaauw would not so-recognize. Returning to the Lyme disease example, if \( P \)’s medical record states incorrectly that \( P \) has Lyme disease, \( Q \)’s reasonable inference would be false. \( Q \) would believe that \( P \) has Lyme disease and \( Q \) would be justified in that belief, but the belief would be false. \( Q \) therefore does not know that \( P \) has Lyme disease (for one cannot know something that is false). On a knowledge account of privacy, such that \( P \)’s privacy regarding \( O \) with respect to \( Q \) decreases only if \( Q \) gains knowledge of \( P \) regarding \( O \), \( P \)’s privacy regarding his health status with respect to \( Q \) would not decrease.

What is important, though, is that despite this disagreement about the particular nature of the privacy relation, both accounts understand privacy as involving a three-part, or \( POQ \), relation. More strongly, understanding privacy as involving a three-part relation is compatible with any plausible account of the nature of the privacy relation. Consider two of the predominant views of privacy in the literature: first, that privacy is fundamentally about access to information, and second, that privacy is about control of information. On access accounts, privacy turns on whether others physically access, cognitively access, or have the ability to physically or cognitively access one’s information. Thus, on access accounts, a person’s privacy does not depend on whether one has the ability to prevent others from impinging her privacy. On control views, one’s having privacy depends on whether one has the ability to decide who can access information about her. So, one can lose privacy if information about her is dispersed (and hence out of her control), even if others do not or cannot actually access that information. Notice, though, that on either type of view, we can articulate some person \( P \), some domain of information \( O \), and some person or persons with respect to whom \( P \) has privacy regarding \( O \). On access views, \( P \) will have privacy regarding \( O \) with respect to \( Q \) if \( Q \)’s access to \( O \) regarding \( P \) is limited in the relevant way. On control views, \( P \) will have privacy regarding \( O \) with respect to \( Q \) if \( P \) has the power to control whether \( Q \) can access information in \( O \) regarding \( P \).

Making Comparisons

Fixing a means of denoting privacy relations \( (POQ) \) allows us to describe one privacy state in isolation. However, there are two problems. First, it is crucial for understanding privacy and context to be able to compare privacy across, for example, time, technology, place, and other relevant variables. Once researchers can make those comparisons with some precision, empirical social scientists can begin to account for the causes of any differences. Second, there are different conceptions of privacy (e.g., control, knowledge, access), all of which we want to be able to compare. That is, we do not want to tie this model to any particular conception of the nature of privacy.

In order to accommodate this last problem, we can use terms to represent particular conceptions of privacy that might obtain in any \( POQ \) relation. Hence, let \( \alpha \) represent an access account of privacy, and \( \alpha_{POQ} \) represent a particular three-part privacy relation under that conception. Table 1 shows a standardized, but non-exclusive, set of terms to refer to four principal conceptions of privacy.

---

1 Examples of access accounts include (Powers 1996; Allen 1988, 15; Gavison 1984, 349–50; Parent 1983, 269).
2 Examples of control accounts include (Moore 2010; Westin 1967, 7; Rachels 1975).
Table 1

<table>
<thead>
<tr>
<th>Conception of Privacy</th>
<th>Associated Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td>( \alpha )</td>
</tr>
<tr>
<td>Control</td>
<td>( \kappa )</td>
</tr>
<tr>
<td>Particularized Judgment</td>
<td>( \pi )</td>
</tr>
<tr>
<td>Knowledge</td>
<td>( \nu )</td>
</tr>
</tbody>
</table>

Consider the case of \( P \)'s privacy regarding his health information. We might want to compare \( P \)'s privacy in that regard with respect to various entities. So, \( P \) likely has relatively little privacy regarding his health information with respect to his doctor, but he might have more privacy in this regard with respect to his neighbor. To represent this difference we will need to expand our formula. Let \( P \) denote Peter, \( O_1 \) denote medical information, \( Q_1 \) denote Peter’s doctor, and \( Q_2 \) denote Peter’s neighbor. In the normal case, the following will be true:

\[
\alpha_{P_0,Q_1} < \alpha_{P_0,Q_2}
\]

That is, Peter’s privacy regarding his medical information with respect to his doctor will be less than Peter’s privacy regarding his medical information with respect to his neighbor. Now, let \( O_2 \) denote Peter’s gardening habits. The following will be true:

\[
\alpha_{P_0,Q_1} > \alpha_{P_0,Q_2}
\]

That is, Peter’s privacy regarding his gardening habits with respect to his neighbor will be less than Peter’s gardening habits with respect to his doctor.

However, if Peter’s doctor is the same person as Peter’s neighbor, then:

\[
\alpha_{P_0,Q_1} = \alpha_{P_0,Q_2}
\]

And:

\[
\alpha_{P_0,Q_1} = \alpha_{P_0,Q_2}
\]

This example simply analyzes a single subject (\( P \)) across different domains (\( O \)) and third parties (\( Q \)). The framework, though, helps us describe privacy relations according to variables such as time and location. Consider, for example, records of persons’ real property. In the U.S. municipalities’ real property records are public records and anyone may access those records. Prior to the digitization of those records, the uptake of the Internet, and the move to place public records online, one generally had to make a request by mail, by fax, or in person to receive those records, and one generally had to pay for processing, photocopying, and postage. Now, in many places one can simply enter a person’s name, a property address, or a parcel number in an online form and receive property records immediately and for free. We can represent this difference using our framework.

Let \( P \) represent a property owner in Greenacre, a municipality in the U.S. Let \( O \) represent information about real property (tax assessment value, property description, purchase price, encumbrances, and so forth). Let \( Q \) represent the general public. Suppose that in the pre-Internet era (\( T_1 \)) Greenacre kept its property records in paper files, which could be accessed in person at City Hall during standard business hours, for a standard fee. However, as of 2010 (\( T_2 \)) Greenacre keeps all of its property records in an electronic database, which may be accessed by members of the public on the city’s website free of charge.

We can easily see, in Table 2, the relation between privacy in Greenacre before and after the database.
Table 2
Privacy state comparison across time

<table>
<thead>
<tr>
<th>Variable</th>
<th>$T_1$</th>
<th>$T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy relation</td>
<td>$\alpha_{POQ}^T &gt; \alpha_{POQ}$</td>
<td></td>
</tr>
</tbody>
</table>

We can replace the table by modifying the notation, including not only the privacy relation $\alpha_{POQ}$, but also adding the relevant variable, in this case time ($T_1$ and $T_2$). Hence, we can represent the overall privacy relation of a property owner regarding information about her property with respect to the general public as follows:

$$\alpha_{POQ}^{T_1} > \alpha_{POQ}^{T_2}$$

We can also construe the change at Greenacre as a change in technology rather than as a change in time. That is, we can analyze it as the difference between paper-based records and digitized, online records, represented slightly differently:

$$\alpha_{POQ}^{Paper} > \alpha_{POQ}^{Digital}$$

Indeed, a similar notation can be used to represent whatever comparison one wishes to make. So, rather than comparing privacy in Greenacre over time or across technologies, we might instead wish to compare privacy regarding property information between Greenacre and Blueacre. If Blueacre, even at this late date, has not created an Internet-accessible electronic database of its property records, the following would represent property owner privacy in the two locales:

$$\alpha_{POQ}^{Blueacre} > \alpha_{POQ}^{Greenacre}$$

We can also combine them:

$$\alpha_{POQ}^{Blueacre \cdot T_1} = \alpha_{POQ}^{Greenacre \cdot T_1}$$

Whereas:

$$\alpha_{POQ}^{Blueacre \cdot T_2} > \alpha_{POQ}^{Greenacre \cdot T_2}$$

Hence, the framework here specifies and isolates aspects of privacy’s context, and is flexible enough to account for different variations. These include, but are not limited to, place, time, and technological developments.

**Toward Social Scientific Explanation of Variance**

This type of structured, rigorous analysis encourages us to look at privacy relations within specific situations: in particular places or times, or under various technological conditions. Once this work is complete, social scientists – and indeed all those who seek to determine the reasons behind the variations elicited – can treat the resulting privacy comparisons as bases for further research. For instance, once the privacy comparison related to presidential campaigns in the US and France presented above is established, scholars can seek to explain the causal factors behind the differences. It may well be that France’s political culture has been so influenced by the presence of a centralized, powerful state that its government is more likely to demand a fuller accounting of donations from its candidates. By comparison, Americans’ tendency toward skepticism, antagonism toward state action, or affinity for small-scale political actors may contribute to the exclusion of sub-$200 donations from federal reporting requirements.

We can also imagine any number of further applications and comparisons, depending on the interest of the social scientist or other analyst. Perhaps computer scientists, information scientists, and designers of technological systems will wish to evaluate the privacy impacts of existing technologies with an eye toward predicting privacy outcomes of future technologies (see, e.g., Detweiler et al. 2011). Or
sociologists may wish to look into the past to compare privacy states under various social regimes in order to make predictions about societal outcomes. Others may wish to examine privacy regimes in various historical periods to identify causes of differences in privacy protections.

**Conclusion**

Our task here has been to advance privacy inquiry by providing a bridge between several discrete areas of privacy scholarship: work emphasizing the importance of privacy's context, philosophical work regarding conceptions of privacy, and empirical social science looking at differences in privacy regimes and underlying causes of such difference. To do so we’ve offered a framework and conceptual foundation for isolating aspects of privacy and privacy's context and for comparing other aspects of privacy's context: privacy is a three-part relation between some person, persons, or entity $P$, some set of propositions or domain of information $O$, and some other person, persons, or entity $Q$ with respect to whom $P$ has privacy regarding $O$.

We have argued that the framework is important insofar as it forces specification regarding these three necessary aspects of privacy and, hence, allows for comparing privacy across contexts such as time, location, and polity. Although it forces some specificity, it is flexible insofar as it allows comparisons of myriad contexts and accommodates different philosophical conceptions of the nature of privacy (access, control, knowledge, inference) and types of privacy protections (legal, moral, technological).

**References**


Service Learning in Action: Integrating Reflection to Deepen the Educational Experience

Benjamin K. Baron
University of North Texas
College of Information
Benjamin.Baron@unt.edu

Barbara Schultz-Jones
University of North Texas
College of Information
Barbara.Schultz-Jones@unt.edu

Abstract

A constructivist learning environment perspective was applied to the potential for library science international service learning projects to provide a short term learning environment that introduces pedagogy, provides the opportunity to develop and apply new skills, energizes students to develop cultural competence, and develops reflective skills towards the pursuit of wisdom. Through the service learning experience, students are able to contribute directly to underserved communities while evolving their own cultural intelligence. The role of reflection in the service learning environment emerged as the dimension with highest significance in terms of short term educational goals of the program and long term student development as citizens of the world. Students reported that the actual experience of reflective behavior exceeded their anticipated levels of reflection at the beginning of the project. Professors reflected on the overall results in order to enrich the experience for future participants.

Keywords: learning environments, international service learning, assessment, reflection, pedagogy

Introduction

Education ideally fulfills two outcomes: it empowers individuals with the knowledge necessary to approach economic challenges from a position of strength, and it provides richness of experience by exposing the learner to facets of culture that allows for the informed and responsible use of gained knowledge. These objectives can be met by service learning projects (SLP) that provide a temporary learning environment to accomplish educational goals.

Since 2003 the College of Information (COI) at the University of North Texas (UNT) has sponsored library science SLPs in Thailand (2003 to 2006), Albania (2008), Ukraine (2010), Peru (2011), and Russia (2012). In each setting, students and faculty worked for a three week period as a project team with partner educational institutions to assess and improve library service for K-12 international students. As with any SLP, previous projects have provided students from UNT with a combination of real-world practice in school library settings along with exposure to cultural experiences that cannot be replicated in the classroom setting.

For this study, "service learning" is defined as a course of study that integrates classroom based instruction and hands-on activities with exposure to local concerns and cultural influences, ultimately directed toward meeting local needs (Florman, Just, Naka, Peterson & Seaba, 2009). Service learning differs from volunteerism in that it "deliberately integrates community service activities with educational objectives" (Bringle & Hatcher, 1999, p. 179). "International", for the purposes of this study, is defined as occurring in a country outside of one’s country of residence or origin. In the case of this study, all of the participants are citizens of the United States, with the project occurring in Russia.

Acknowledgements: We would like to thank the Anglo-American School of Moscow for its generosity and the opportunity for the University of North Texas, College of Information School Library study abroad participants to improve library services at the AAS Moscow libraries while learning valuable, real world lessons in providing library services, teamwork, and project execution. In particular we would like to thank John Bishop for his enthusiasm, assistance, and guidance. Benjamin would also like to thank Claudia Wayland for her service as both editor and motivator.
Copyright is held by the authors.
Service learning projects (SLPs) provide students of all ages and backgrounds with opportunities to integrate real world experiences into the learning framework. The questions this study seeks to consider are: how do service learning environments (SLEs) serve the education of undergraduate and graduate library science students in terms of core competencies and cultural development? Furthermore, how can these experiences benefit underserved communities and cultures, specifically in countries outside the United States?

Numerous studies detail the educational and experiential value of academic SLPs in both local and international settings (Barry, 2011; Bringle & Hatcher, 1999; Burns, 1998; Giles & Eyler, 1994; Nutefall, 2009; Riddle, 2003); furthermore, each of these studies highlights the importance of reflection as a differentiating factor between volunteer projects and SLPs. Reflection is an attribution that informs and transforms knowledge and action (Risko, Roskos & Vukelich, 2002). By creating learning opportunities where students can reflect and assess their own learning the educational value of a service learning environment can be extended to support the development of reflective thinkers (Jones & Dotson, 2010).

While there are a limited number of studies that detail methodologies for assessing the impact of service learning on students, development of a comprehensive toolset to assess the effectiveness of SLPs as learning environments is necessary. This pilot study seeks to apply a constructivist learning environment perspective to the assessment of an international SLP involving Library and Information Science students.

### Theoretical Framework

Theoretically, the study was grounded on the constructivist point of view that learning environments are individual “constructions” (Tobin & Fraser, 1998, p. 626) which are neither independent from, nor external to their participants (Lorsbach & Basolo, 1998). Lorsbach and Basolo state that students and teacher simultaneously contribute to the creation of their learning environment; they interact within it and individually perceive it as observers. This view emphasizes the importance of student and teacher perceptions in the study of learning environments. These perceptions reflect one’s personal interpretation and beliefs of the degree to which the physical and social setting of a class permits or hinders learning.

The aspect of learning environment research which focuses on conceptualizing, assessing, and investigating how classroom participants perceive the socio/psychological aspects of their learning environment has received significant attention over the past 30 years and has been used to evaluate educational reforms (Dryden & Fraser, 1998; Taylor, Fraser & Fisher, 1997). Assessing how school library participants perceive the socio/psychological aspect of their learning environment has been studied recently (Schultz-Jones & Ledbetter, 2009, 2010a, 2010b) and draws on the foundation of learning environment research established for the classroom, with emphasis on the science classroom.

The context for learning environment research uses the following definition: “Learning environment refers to the social, psychological and pedagogical contexts in which learning occurs and which affect student achievement and attitudes” (Fraser, 1998a, p. 3). The consideration of a classroom’s learning environment built on the foundational work of Lewin (1936) and Murray (1938) and advanced with the development of evaluation instruments. As research on classroom learning environments continued to grow, the number of evaluation instruments increased in Western and non-Western countries (Fraser, 2002, 2007). These instruments have been used in several research studies reviewed by Fraser (1998b; 1998c), including investigations of associations between learning outcomes and classroom environments (McRobbie & Fraser, 1993) and the evaluation of educational innovations (Ogbuehi & Fraser, 2007; Maor & Fraser, 1996; Martin-Dunlop & Fraser, 2008; Monsen & Frederickson, 2004). Despite these documented efforts an examination of the context within which learning occurs in a project team school service learning environment has not included an application of the learning environment paradigm for assessment. Prior to 2009, these instruments had not been used in a school library setting (Schultz-Jones & Ledbetter, 2009, 2010a, 2010b).

For the purposes of this study two learning environment instruments were selected to form the basis of assessment because they are based on a psychological view of learning that focused on students as co-constructors of their own knowledge (Goh, Young & Fraser, 1995; Fraser & O’Brien, 1985). One has been used in past school learning environment research and one was recently developed for the school library environment. The questionnaires have two distinct applications: student assessment of the preferred learning environment and student assessment of the current learning environment. Both
instruments reflect the classification of scales according to Moos’ (1974) scheme for classifying human environments into the three basic dimensions of Relationship (the nature and intensity of personal relationships within the environment, the extent to which people are involved in the environment and support and help each other), Personal Development (basic directions along which personal growth and self-enhancement tend to occur), and System Maintenance and System Change (extent to which the environment is orderly, clear in expectations, maintains control and is responsive to change) (Fraser, 1998c).

The My Classroom Inventory (MCI) is a valid and reliable instrument for the assessment of students’ perceptions of constructivist classroom learning environments. While it was developed for use at the primary school level for children aged 8-12 (Fraser, 1998c), the MCI was selected as a basis for use with this assessment of a project team because of its distinctive ability to characterize the specific dimensions of satisfaction, competition, friction, difficulty and cohesion. These five dimensions relate to and support the development of inquiry based learning, and typifies the experiences of project team members as evidenced by prior service learning projects.

The How My Library Supports Inquiry (HMLSI) questionnaires were developed to evaluate student perceptions of student learning in an inquiry-based school library learning environment and the effect of this environment on student literacy and by extension, the social good (Schultz-Jones & Ledbetter, 2013). The development of these instruments was based on the MCI and the What Is Happening In My Class (WHIC), developed by Fraser, Fisher and McRobbie (1996) for use in the context of secondary school level classroom environments and introduced the dimension of “reflection”. Each HMLSI questionnaire uses 28 items comprising seven scales to measure students’ perceptions of the degree to which certain psychosocial factors are prevalent in the school library: reflection, librarian support, involvement, investigation, task orientation, cooperation, and equity. These dimensions relate to and support the development of inquiry based learning and a constructivist learning environment. The factor structure, internal consistency reliability, discriminant validity, and the ability to distinguish between different classes and groups were supported for both instruments. Validation of these instruments used data from a Texas study of 872 elementary students and 639 secondary students, principal components factor analysis with varimax rotation and Kaiser Normalization to confirm the a priori structure of the questionnaires.

The MCI and HMLSI instruments were further adapted to accommodate data collection in a temporary project team learning environment at an international school setting by adapting the language to recognize the service learning environment. The modified instruments are named My Perception of Service Learning Environments (MPSLE) with ten elements assessed: Reflection, Faculty Support, Involvement, Investigation, Task Orientation, Cooperation, Equity, Satisfaction, Friction, and Cohesion. Incorporation of these instruments contributes a unique design for consideration of a variety of school library contexts within the burgeoning field of learning environments research (Nix, Ledbetter, & Fraser, 2001).

Since the school library provides a learning environment for the development of inquiry skills to advance student achievement, regardless of national location, the application of these assessment tools to an international service learning project in a school library context is worth examination. Further, this study of psychosocial aspects of the learning environment offers potentially valuable ideas for incorporation by library science educators and by extension the students with whom they interact. Much of the learning environment research focuses on the perception of a learning environment and the extent to which this perception matches what is preferred by students. There is a gap in the literature, however, on efforts to use this research as an assessment tool that can be used to transform a temporary learning environment to meet and enrich educational objectives. With this research study we propose to move the focus on perception to transformation with service learning projects by emphasizing the role of reflection in the learning process.

The concept of reflection was introduced by Dewey (1933) and was considered to be a cognitive process of problem solving or thinking that was used to resolve an issue (Dewey, 1944). As Hatton and Smith (1995) record, he viewed reflection as “an active and deliberative cognitive process, involving sequences of interconnected ideas which take account of underlying beliefs and knowledge” (p. 34). Dewey (1933) and Schon (1983, 1987) also consider reflection as a cyclical approach that integrates problem identification, contextualizing the problem, data gathering and consideration of possibilities. This process applies not only to activities associated with a project team but individually and collectively as students make sense of a new cultural environment. And the value of reflection is the extension of
inquiry-based learning to incorporate lifelong problem solving skills and the ability to make meaning of experiences.

Furthermore, developing a well-rounded, cosmopolitan point of view will serve the information professional in his or her life of service, regardless of location or time (Monteil-Overall, 2010). In particular, librarians of all varieties will face cross-cultural hurdles that will test their understanding and ability to interact with students or patrons. Participation in a SLP is one method of expanding one’s perceptions and providing opportunities for reflection on the experience is a constructive way to deepen and extend the educational experience.

Research Setting

In May 2012 twenty students and five faculty members with the UNT-COI service learning study abroad project traveled to Moscow and Saint Petersburg, Russia to assist a large K-12 International Baccalaureate accredited school. The school delivers educational services with and to a diverse population, with 2,000 students and 40 teachers from more than 60 countries. Two school libraries reside within the Moscow and St. Petersburg schools, one for the primary and middle grades and one for the high school students. The project involved developing a recommendation for space appropriation for a consolidated school library in the Moscow school, organizing the school libraries and developing the associated and appropriate library policies that support the school library media center and the school curriculum.

The project encompassed common library practices (cataloging, policy development, circulation, etc.), cultural exposure experiences (visiting museums, locations of interest, and attending cultural events such as concerts and ballet performances), and cultural interaction opportunities (round table meetings with members of the Russian Association of School Librarians in Moscow and St. Petersburg). The learning environment, therefore, extended beyond the school boundaries to encompass a broader cultural context.

The group consisted of two undergraduates, fifteen master’s level students, one doctoral dissertation candidate, and one non-degree seeking student who graduated from the UNT Masters of Library Science program in 2009. Of the five faculty members, one was the Dean of the College who accompanied the group for the introduction of the project only. The remaining four faculty members provided instruction and support to the student project teams. The project team was further divided into five task teams, each with specific goals: technology support, policy development, catalog assessment and revision, organization and cataloging of classroom sets, and space planning. Each task team took primary responsibility for the task team goal and each task team member rotated among all task teams so everyone had experience with each task. The teams were located in a large project area where communication was easily accommodated and all team’s had convenient access to each other and the accompanying faculty members.

Both the participants from UNT and the sponsoring organizations used English as the primary language for communication at the project site; the majority of interaction with native Russian speakers was isolated to experiences outside of the library setting.

Methodology

In order to evaluate and improve the project for future participants, nineteen of the students provided feedback related to the SLP learning environment. Student participants were provided with a total of four data collection instruments: The BFI-46A Five Factor Model personality survey, a student-investigator generated social network analysis (SNA) survey, and the My Perception of Service Learning Environments (MPSLE) surveys based on modified versions of the How My Library Supports Inquiry (HMLSI) (Schultz-Jones & Ledbetter, 2012) and the My Classroom Inventory (MCI). The initial set of surveys (MPLSE-Preferred and BFI-46A) was distributed by the student investigator at the beginning of the project. Participants responded to surveys on their own time before depositing completed surveys in an envelope which was sealed at the conclusion of data collection. The same procedure was used for the final set of surveys (MPLSE-Actual and SNA questionnaire). Collected data remained sealed until coding and data analysis was conducted.

Students also created and submitted a photo journal based on guiding questions as part of the reflective component of the SLP and completed a team participation evaluation instrument directed at
feedback for each individual team member. This report will focus only on the learning environment surveys. Example questions from the modified instrument are provided in Table 1.

### Table 1

*Instrument Example From “My Perception of Service Learning Environments” (MPLSE) Survey*

<table>
<thead>
<tr>
<th>Preferred</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supervising Faculty Support</strong></td>
<td></td>
</tr>
<tr>
<td>The supervising faculty would discuss project activities with me.</td>
<td>The supervising faculty discussed project activities with me.</td>
</tr>
<tr>
<td>The supervising faculty would be interested in my project questions.</td>
<td>The supervising faculty was interested in my project questions.</td>
</tr>
<tr>
<td>The supervising faculty would move about the project teams to talk with me.</td>
<td>The supervising faculty moved about the project teams to talk with me.</td>
</tr>
<tr>
<td>The supervising faculty’s questions would help me to understand what I am contributing to the project.</td>
<td>The supervising faculty’s questions helped me understand what I contributed to the project.</td>
</tr>
<tr>
<td><strong>Friction</strong></td>
<td></td>
</tr>
<tr>
<td>Conflict would exist between team members at the project site.</td>
<td>Conflict existed between team members at the project site.</td>
</tr>
<tr>
<td>Some team members at the project site would be aggressive.</td>
<td>Some team members at the project site were aggressive.</td>
</tr>
<tr>
<td>Certain team members would always want to have their own way at the project site.</td>
<td>Certain team members always wanted to have their own way at the project site.</td>
</tr>
<tr>
<td>Team members at the project site would often engage in conflict or argue.</td>
<td>Team members at the project site often argued needlessly.</td>
</tr>
<tr>
<td><strong>Reflection</strong></td>
<td></td>
</tr>
<tr>
<td>I would like to think about the overall consequences of the project.</td>
<td>I thought about the overall consequences of the project.</td>
</tr>
<tr>
<td>New learning would relate to any questions I ask inside and outside of the project site.</td>
<td>I acquired new learning related to questions I asked inside and outside of the project site.</td>
</tr>
<tr>
<td>I would learn how team work is a part of life both at and away from the project site.</td>
<td>I learned how team work is a part of life both at and away from the project site.</td>
</tr>
<tr>
<td>I would acquire skills and knowledge that apply to the world inside and outside of the project site.</td>
<td>I acquired skills and knowledge that apply to the world inside and outside of the project site.</td>
</tr>
</tbody>
</table>

Students answered four, five point Likert scaled questions for each of ten dimensions: Reflection, Faculty Support, Involvement, Investigation, Task Orientation, Cooperation, Equity, Satisfaction, Friction, and Cohesion. Each student provided separate responses about the learning environment at two instances in time: the first set of responses measured preferences at the beginning of the service learning project, while the second set of responses measured feedback for the same ten dimensions based on
actual experiences at the end of the project. Data collection took place over a three week window. Data was analyzed using a standard t-test comparing means for each dimension.

Findings

Given the relative homogeneity of the group (the majority of the group was Caucasian females between the ages of 25 and 60), no analyses were conducted on sub-groups defined by ethnic, gender, or academic program based characteristics within the overall participant group. Table 2 provides results for all scales.

Table 2
All participants’ Preferred and Actual scales compared using paired sample t-test

<table>
<thead>
<tr>
<th>Scales</th>
<th>Preferred</th>
<th>Actual</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reflection</td>
<td>4.12</td>
<td>4.49</td>
<td>.23</td>
</tr>
<tr>
<td>Faculty Support</td>
<td>4.61</td>
<td>3.99</td>
<td>.22</td>
</tr>
<tr>
<td>Involvement</td>
<td>4.33</td>
<td>4.25</td>
<td>.31</td>
</tr>
<tr>
<td>Investigation</td>
<td>4.16</td>
<td>4.16</td>
<td>.36</td>
</tr>
<tr>
<td>Task Orientation</td>
<td>4.42</td>
<td>4.51</td>
<td>.21</td>
</tr>
<tr>
<td>Cooperation</td>
<td>4.54</td>
<td>4.58</td>
<td>.34</td>
</tr>
<tr>
<td>Equity</td>
<td>4.62</td>
<td>4.47</td>
<td>.27</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>4.36</td>
<td>4.30</td>
<td>.38</td>
</tr>
<tr>
<td>Friction</td>
<td>1.95</td>
<td>2.87</td>
<td>.96</td>
</tr>
<tr>
<td>Cohesion</td>
<td>4.30</td>
<td>4.26</td>
<td>.43</td>
</tr>
</tbody>
</table>

df = 18
*α = 0.05

Of the ten dimensions that were examined in this study, three show significance: Reflection (t-stat = 2.22, t-crit 2.10, α = 0.05), Faculty Support (t-stat = -4.23, t-crit 2.10, α = 0.05), and Friction (t-stat = 3.02, t-crit 2.10, α = 0.05). These dimensions are discussed in the following section.

Discussion

The role of experience in the education process is well established by both John Dewey (1938) and David Kolb (1984), who felt that direct experience resulted in a transformative process when coupled with instruction. In the spirit of this perspective, reflective behavior rises to prominence. Reflection is an essential component of the service learning model that distinguishes it from the classroom environment. Bringle and Hatcher (1999) point out that, “[service learning] students frequently encounter new circumstances and challenges. These experiences often create dissonance, doubt, and confusion” (p. 180). Consequentially, reflection allows students to reconcile differences between beliefs, perceptions, and experiences. In the case of the Russia 2012 project, students reported that the actual experience of reflective behavior exceeded their anticipated levels of reflection at the beginning of the project. A number of stresses may have contributed to this finding, among them: an unfamiliar spoken and written language, library practices and tasks which may have been new to certain students, and social and cultural practices that may differ significantly from those practiced at home. Moreover, the ad hoc nature of the social group, comprised of near strangers, may also contribute to this reflective behavior as participants evaluate new social connections and interactions within the context of the project and the experience.

Qualitative data, in the form of individual photo journals with answers to a variety of challenge questions, will be assessed to glean the focus of student attention since the strength of Reflection as a learning environment dimension supports further focus on this element for future project iterations. Use of this data collection tool along with directed prompts or questions may provide researchers with broader perspectives not captured through the relatively limited responses offered by survey instruments.

Scores for Faculty Support indicate that students did not receive levels of faculty support as expected. When the correlation between the differences in perceptions of preferred and actual faculty support and the differences in perceptions of preferred and actual reflective experience is considered, a
positive relationship emerges (0.33, df=36, α=0.05), indicating that as faculty support increases, reflective behavior may increase. When actual faculty support is correlated with actual reflection, an even stronger result emerges (0.64, df=36, α=0.05), indicating that increasing faculty support may be a key contributing factor for a more reflective learning environment. Given the nature of the project, independent work skills and the ability of participants to make decisions is assumed, but younger generations of student may crave more interaction with peers and faculty (Becker, 2009). Conversely, the assessment encompasses four faculty members while the discrepancy may be directed at fewer than four of this faculty. Regardless, this dimension requires significant and continued educator reflection and raises the following questions: in what areas did faculty support exceed or fail student expectations, and how can faculty further support reflective behavior to encourage and deepen educational practices? This dimension will be tracked in future projects following modifications of behaviors and practices.

Friction also emerged as a dimension of significant interest. Friction displayed the greatest amount of variance between actual (2.20) and preferred levels (0.96), exceeding all other variances for actual (nearly 5:1) and preferred (between 2:1 and 3:1) dimensions. Additional analysis of the data shows that increased actual perceptions of friction were expressed by the majority of the participants. Of the 19 participants, 13 (68.4%) reported greater actual levels of friction than expressed in the preferred data. Of the remaining six participants, four participants (21%) reported less actual friction than preferred, and the remaining two participants (10.5%) reported no difference in their preferred and actual perceptions of friction. When the correlation between the differences in perceptions of preferred and actual friction and the differences in perceptions of preferred and actual reflective experience is considered, a negative relationship emerges (-0.39, df=36, α=0.05), indicating that as friction increases, reflective behavior may suffer. When actual friction is correlated with actual reflection, a similar result emerges (-0.33, df=36, α=0.05), indicating that limiting or preventing friction may contribute to a more reflective learning environment. Further analysis of the team participation evaluation instrument may also provide insight into the results for this dimension. Future projects will provide the opportunity to see if friction is unique to this participant group or is a common feature of service learning environments.

While both Faculty Support and Friction displayed significant results, neither has received the same treatment in the literature as Reflection, indicating opportunities for further exploration. Consideration of this dimension and examination of additional ways to provide reflective opportunities, either guided or interactive, will be explored to further enhance the educational opportunities for students involved in future service learning projects. Further analysis and consideration of Faculty Support and Friction will also contribute to building educationally rich and fulfilling service learning projects in the coming years.

**Conclusion**

Assessing the learning environment of service learning projects is limited to this current study. Regardless of this limitation, the results are encouraging and provide directions for improvement and transformation of the service learning project experience. The K-12 International Baccalaureate accredited school of Moscow and St. Petersburg has invited UNT to continue the efforts of the project team in two subsequent years. This provides the opportunity to evolve and transform the project learning environment as we incorporate changes and measure the results of our efforts.

The use of additional assessment tools must be considered for future iterations of the project. These tools might include offering additional focus questions (some of which may request the use of narration free images as an expression of a participant’s perceptions), speculative questions that explore the “whys” and “what” of participant thinking, and daily journaling activities to determine if reflective behavior is ongoing and builds upon previous experiences. By building a comprehensive bank of data related to reflection, researchers can determine if particular activities, be they cultural, educational or social, are more closely associated with reflective behavior. In addition, considering reflective behavior in the context of personality traits will allow researchers to tailor tasks and experience opportunities to maximize the potential for participant reflection. Moreover, assessment of this reflective behavior should not be isolated to one point in time and ongoing contact with participants both past and future will be established.

The cyclical approach to reflective thinking will become evident as we examine the practice of the service learning environment and emphasize relevant student self-assessment in learning, knowing and reflecting within a culturally diverse environment. However, for this to occur, researchers and faculty are
obligated to integrate this approach within future service learning projects and consider longitudinal evidence of the impact of modifications. In this way a progression of enriched experiences should be established.

References


innovative science course designed for prospective elementary teachers. *International Journal of 
Science and Mathematics Education, 6*(1), 163-190.

McRobbie, C. J., & Fraser, B. J. (1993). Associations between student outcomes and psychosocial 


Monteil-Overall, P. (2010). The effect of service learning on LIS students’ understanding of diversity 
issues related to equity of access. *Journal of Education for Library & Information Science, 51*(4), 
251-266.

Press.


5*(4), 250-261. doi:10.1080/15228950903199271

environment that bridges university field classes and field trips*. Paper presented at the annual 
meeting of the Australian Association for Research in Education, Fremantle, Australia.

Ogbuehi, P. I., & Fraser, B. J. (2007). Learning environment, attitudes and conceptual development 
associated with innovative strategies in middle-school mathematics. *Learning Environments 
Research, 10*, 101–114.

*Journal of Academic Librarianship, 29*(2), 71.

Risko, V., Roskos, K., & Vukelich, C. (2002). Prospective practitioners’ reflection: Strategies, qualities, 
and perceptions in learning to teach reading. *Reading Research and Instruction, 41*(2), 149-175.

Books.


elementary and middle school students’ perceptions*. Research Paper in the Proceedings of the 
13th International Forum on Research in School Librarianship at the 38th International Association 
of School Librarianship Annual Conference, Padua, Italy, September 2-4, 2009.

Schultz-Jones, B., & Ledbetter, C. (2010a). *Assessing school libraries as learning environments: 
Examining students’ perceptions in grades three, four and five*. Research Paper in the 
Proceedings of the 14th International Forum on Research in School Librarianship at the 39th 
International Association of School Librarianship Annual Conference, Brisbane, Australia, 

Schultz-Jones, B., & Ledbetter, C. (2010b). Investigating third through sixth grade students’ perceptions 
of school library learning environments. Research Paper in the Proceedings of the American 

Validation of two new learning environment questionnaires. *Learning Environments Research*.


Tobin, K., & Fraser, B.J. (1998). Qualitative and quantitative landscapes of classroom learning 
environments. In B.J. Fraser & K.G. Tobin (Eds.), *International handbook of science education* 
Understanding Large Scale Online Environments with Qualitative Methods

Dana Rotman
drotman@umd.edu
Yurong He
yrhe@umd.edu
Jennifer Preece
preece@umd.edu
Allison Druin
allisond@umiacs.umd.edu

College of Information Studies, Maryland’s iSchool, Human-Computer Interaction Lab
University of Maryland

Abstract

Large scale online environments, in which there are hundreds of thousands, and often millions of links, present an emerging phenomenon where millions of people come together online to share and consume information. The resulting massive amounts of data have been a fertile ground for quantitative researchers. Qualitative studies of these environments are less common, suggesting that the scale and the constant change of these environments pose considerable challenges for qualitative researchers. We present an exploratory study into the challenges and opportunities experienced by researchers conducting qualitative research in large scale online environments, and a meta-analysis of papers from the ACM Digital Library that reveals how few published research studies use qualitative methods to investigate large-scale online environments. We discuss practical and theoretical issues arising from the unique stance of qualitative researchers in these environments.

Keywords: qualitative research, online research, social media, research methods, communities of practice

Introduction

Large scale online environments, ranging from social networks (e.g., Facebook), to peer production networks (e.g., Wikipedia), information dissemination tools (e.g., Twitter), outlets for creativity (e.g., Flickr, YouTube) and collective curation communities (e.g., Encyclopedia of Life, Pinterest), have become a mainstay of online activity. These environments enable users to exchange information, engage in intellectual production of textual and audio-visual content, collaborate on a massive scale, and engage in various discretionary activities. For many researchers these activities and the traces they leave are a goldmine of research opportunities, enabling an intimate glimpse into the social dynamics of technology use, and the processes of engagement, collaboration, production and attachment on a scale larger than ever before (Online Activities, 2000-2009). Where previously online studies focused on relationships formed in small groups such as dyads, families, communities, and groups within organizations, large scale online environments expand the field to include networked relationships encompassing hundreds of thousands and even millions of people.

Perhaps not surprising, these rich and often relatively easy to collect data drive much of the research done on large scale social environments to be quantitative assessments of the strength and divergence of relations created among users, and the structure of the network created by them (cf. Brandes et al, 2009; Cha et al, 2007; & Paolillo, 2008). Quantitative studies can enable researchers to identify changes and trends, areas of particular activity, information-flow directions, gate-keepers, those at the center of networks or those at the outside, those with many links to others or singletons (cf. Gilbert, 2012; Gleve et al., 2009; Golbeck, 2008; Shamma et al., 2009). Yet quantitative studies are not well suited to answer questions related to human values, motivations and meanings, since they offer a rich but incomplete picture of behavior and intention. We propose that qualitative methods are better suited to answer these types of questions, yet qualitative studies of large scale online environments are relatively rare. Given the rapid development of these environments, the question remains, why have qualitative...
methods not been more widely adopted to study them? A general answer to this question is that these environments present many challenges to doing in-depth qualitative work. Some of these challenges relate to issues of scope, size and the ephemeral structure, tools, and culture of these environments (Rotman et al., 2012).

This paper presents an exploratory study into the opportunities and challenges that qualitative researchers face when studying large scale online environments. To do that we conducted nine interviews with researchers known for doing qualitative research, in which we discussed the ways in which researchers reconciled the opportunities offered by large scale online environments and the practical and theoretical problems they face. In addition we examined papers from the ACM Digital Library in order to gauge the popularity of qualitative methods for studying large scale online environments.

Large Scale Online Environments

These environments comprise websites, social networks and other platforms that enable users to engage in a range of activities on a scale larger than ever before. They can be social or egocentric networks like Facebook, collaborative tools like Wikipedia, content production repositories like YouTube and Flickr, or any number of tools and interfaces that are not bound by a specific domain, locale, or size. Based on a networked structure of content and various types of interactions, these environments grow continuously, although some may eventually fail. The online activity produces masses of data detailing user habits from the mundane to the extraordinary.

The massive data created, shared and consumed in these environments are sometimes referred to as “Big Data.” Big Data presents “the big picture from the minutia of our digital lives” (Fisher et al, 2012). The common definitions of Big Data emphasize the size of the data, or its relationality (“Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data” (Boyd & Crawford, 2012)) and do not refer to the other properties of the environment in which this data is situated. The concept of Big Data is useful for analyses focused solely on the data that can be derived from online interaction, but is lacking in terms of comprehensiveness and understanding the underlying processes, human values, motivations, and meanings that are associated with these data. For this reason, we prefer to use the term “large scale online environments” which incorporates the data, the tools that are used to create, share, and consume it, and the overall interaction around it.

Qualitative Research of Large Scale Online Environments

Qualitative research has worked its way to acceptance in some areas of information science, Human Computer Interaction, Computer Supported Cooperative Work, and social computing. Discussing the rich, broad, and often contentious history of qualitative work in these fields is beyond the scope of this paper (and is detailed at length elsewhere (Adams et al., 2008)), but we will try to highlight pivotal points that offer a backdrop against which we situate our research.

Qualitative methods vary immensely: from ethnography and ethnographically-inspired methods (e.g. rapid ethnography and detailed case studies), critical studies and phenomenology, to contextual design and “in the wild” studies. The premise of all these variations is to offer researchers and readers alike a systematic yet naturalistic interpretation leading to an understanding of human behavior within cultures, communities, and technological systems (Patton, 2002). These methods go beyond mere traces of interaction and activity, towards a holistic understanding of implicit concepts of motivations, meanings and attitudes that are sometimes different or divergent from explicit behaviors (Dorr-Bremme, 1985).

Due to its naturalistic and exploratory nature, qualitative work was sometimes at odds with the broad, predominantly positivistic research community. As Barkhuus and Rode (2007) illustrated, until the beginning of the 1990s qualitative work was marginalized and considered lesser than positivistic evaluations. But even after the 1990s, qualitative research, and particularly ethnography, was not embraced without debate (Dourish, 2006). To gain acceptance into the broader research community, qualitative methods were sometimes modified: from long immersion in the field in cultural anthropology (ranging from months to years) rose “rapid ethnography” (Millen, 2000), a compressed, short-term form of ethnographic inquiry. Other methodological variations, highlighting qualitative principles (i.e., observations, interviews, and working “in the wild”) have also been accepted.
On a different trajectory, during the past two decades online environments became fertile grounds for qualitative research. Online ethnography, for example (Garcia et al., 2009; Hine, 2000) adopts principles of ethnographies formed in offline environments, and applies them to virtual environments, with the necessary adjustments (e.g. distant observation, extensive use of pre-formed textual and visual artifacts such as transaction logs, conducting interviews using mediated technology). Similar adaptations of qualitative methods are used to study virtual worlds (Boellstorff et al., 2012). The strength of online qualitative work lies in providing comprehensive descriptions of structures, interactions and content. Yet, despite the growing acceptance of qualitative methods, only a small portion of qualitative studies have departed from small scale systems and communities, and have been conducted in large scale environments (cf. Baym, 2000; Boyd, 2007; Nardi, 2010).

**Methods**

To gain an understanding of how qualitative methods are practiced in current research of large scale online environments, we took a two-step approach: first, we interviewed leading researchers who do qualitative work in large scale online environments. We then used one sub-genre of qualitative methods (“ethnographically inspired” - methods that are adapted from and resemble aspects of accepted ethnographic practices, but do not necessarily adhere to ethnography in its original anthropological sense) and conducted a qualitative meta-analysis (Altheide, 1987) of relevant papers published in leading Special Interest Group in Computer Human Interaction (SIGCHI) conferences over the past 25 years. We conducted this second step in order to gain an overview of the extent to which large scale online environments are studied using qualitative methods either solely or to complement the use of other methods.

**Interviews with Qualitative Researchers**

Recognizing that this study aims to capture the experiences and challenges that researchers face when conducting qualitative work within large scale online settings, we started by conducting nine in-depth interviews with leading researchers who identified themselves as having conducted qualitative research and are known for doing qualitative research in large scale online environments. The interviewees were chosen based on their methodological choices as reflected in their publications in related journals, books, and conferences. We tried to cast a wide net when inviting interviewees to participate in our study, yet many researchers declined and noted that they only did qualitative research in small scale online settings or offline. The final group of nine interviewees ranged across PhD candidates (2) at the later stages of their studies, to tenure-track (2), and tenured faculty or senior researchers in industry (5), providing a relatively broad scope of positions and practices. Though all were semi-structured interviews focusing on the broad theme of doing qualitative research in large scale online environments, the direction each interview took was shaped by the individual interviewee’s experience and thoughts and the dialog created between the interviewers and the interviewees. Seven interviews were conducted face-to-face and two were conducted over Skype. In five of the interviews, two researchers were present and the other four were conducted by one researcher. The researchers took notes after each interview, and while these were not “field notes” in the observational sense, they were used to compliment the interviews and add dimensions the researchers found to be interesting or important. The interviews were 35-120 minutes long, and were audio recorded. They were later transcribed and coded according to the principles of grounded theory (Strauss & Corbin, 1990), where concepts and themes surfaced from the data itself. In order to protect the privacy of our interviewees identifying details such as name, gender, professional position and affiliation and specific fields of study were removed.

**Meta-analysis of Published Papers from ACM Digital Library**

We used the ACM Digital Library to search for all the papers whose metadata referred to qualitative practices derived from ethnography, such as interviews and participant observation. This query resulted in 664 papers, published from 1988 to 2011.
The papers were first coded in an open, bottom-up, coding scheme, following the principles of grounded theory (Strauss & Corbin, 1990). Two researchers coded a sample of about 8% of the papers, and calculated inter-rater reliability statistics (Cohen’s Kappa) between .42 and .98 for three illustrative codes. Consensus on the codes, where there was disagreement, was reached through discussion. The final codebook was then used to analyze a sample of 311 papers, until conceptual saturation was attained.

Discussion of Findings

We set out to explore the practices, challenges, positions and assumptions of qualitative researchers who study large scale online environments. As is often the case with exploratory work, while we were looking to find the ways in which the researchers faced the challenges presented by large scale online environments, unexpected themes came up from the data. These themes can be defined along two axes: practical and theoretical, which will be discussed below. Although we did not look for differences in junior and senior researchers’ viewpoint, this became apparent from the data: most practical challenges were brought up by the junior researchers, (tenure-track professors and PhD students), although they were later confirmed by the senior researchers; while senior researchers focused on the broader theoretical challenges they’ve faced. This could partially be attributed to the career needs of junior researchers, and to publication demands that require them to engage in hands-on research, while senior researchers face these challenges mainly through their work with other collaborators and with their students. Senior researchers also enjoy the perspective of many years in the field, which may allow them to more freely reflect on theoretical and conceptual aspects of research practices. Figure 1 visually represents the coding scheme and the themes that emerged from the data, separating the two major themes of practical issues and theoretical issues, and highlighting the difference between junior and senior interviewees that presented them.

![Figure 1. Coding themes from the interviews. Practical issues were mostly reflected by junior researchers while theoretical issues were brought up by senior researchers.](image-url)
Practical Issues

Large scale online environments brought practical concerns that affected the way qualitative work is done, these include:

(1) Getting enough data, and getting too much data
(2) Identifying entry points to the field and selecting participants
(3) Ephemerality, interface, and cultural change
(4) Applying ethical oversight

Getting enough data and getting too much data. Interactions in large scale online environments create a tremendous amount of data. The popular assumption is that this data is easily available and can almost be cherry picked by researchers. Our interviewees, however, had a different opinion: while log data is, indeed, relatively easy to obtain, collecting data that is a valid base for qualitative research is not as easy. Issues of accessibility and quality came up often in the interviews. Researchers were frequently faced with missing data, commercial restrictions on data use, or with data that could not be validated or was not usable for qualitative studies. As one interviewee summarized: “We are working with the data that we were able to get… and maybe if you are good you can figure out how to publish it, but mostly you’re just going to know a lot”.

On the other hand, the scope of the data, created a different challenge: qualitative researchers were not only forced to justify their methodological choices (see below), but when acting within the chosen methodology they sometimes felt that they were “drowning in data” or “up to your eyeballs in data” – though these data were not always useful or even relevant. They quickly learned that large scale online environments do not lend themselves well to holistic observation which typifies qualitative studies. The immense scope of the data, comprised of interactions between millions of users, required the researchers to make a-priori decisions about the segments of data that are pertinent for addressing their research questions. As several interviewees mentioned, the sense making process became much more difficult when the quantities of data were outstanding. Size didn’t equal quality, and the data available to them presented just a snapshot of a much larger picture, which was hard to capture and deeply understand.

A different aspect of the data scale was the need to justify methodological choice, which, by nature, ignored most of the data and favored extremely small segments of the total data available:

“If you want to do big data where you have millions of data points, it is really hard to argue that we use these numbers and then we go ahead and we interview 20 people. Because how can they be in any way representative of this big data? So these combinations are really hard to argue for”.

Getting a grounded notion of what users were doing, why they were doing it, and what may be valued and valid research aspects, proved to be difficult.

Identifying entry points to the field and selecting participants. One of the most challenging issues the researchers faced was that of selecting participants within large scale online environments. Traditionally, qualitative studies focused on relatively small groups: that could have been a company or an office, a community, users of a system, etc. Large scale online environments make the selection of the entry point to the field and selecting the relevant unit of analysis immensely difficult. Issues of effective and relevant participant selection came up in all of our interviews. The researchers we studied agonized over the need to scope various types of users and behaviors and finding the right people that will lead them into the field, among the millions of users who may not be as relevant to them. As one interviewee mentioned:

“Effective sampling. How do you do sampling? I’m constantly baffled by that. Things that we did in the past are just not as effective here”

Some of the researchers found ways to confront this problem, by modifying previously used techniques, reaching out to people they know in real life, or casting a wide net when looking for appropriate people and segments of the data:

“Take your big pile of numbers, go through it. Figure out the interesting cases. In some cases figure out if there is an interesting phenomenon taking place and get curious. And then go, because you have a big pile of data, find the indexical people who seem to be doing this a lot, and then find out what they are doing… the ones who are figuring out new ways of doing things. Talk to them. They might be the future. This is not just theoretical sampling … I would say this is actually looking for the biggest performers”
But this approach presented new issues – looking for “the biggest performers” translates into finding the most prominent and talkative users or popular pages, making it “impossible to get or very challenging to get peripheral members. There is just self selection which can be problematic”. The structure of social networks and other collaborative environments does little to enable access to the singleton or small groups that are not highly connected but may present an important phenomenon nonetheless. And while large scale online environments enable researchers to look at all kinds of online behavior, selection will result, in many cases, in data collected from the most convenient areas, and not the most interesting ones.

Ephemerality, interface and cultural change. Large scale online environments constantly change. While changes to the interface can be beneficial to researchers, sometimes allowing serendipitous findings that would have otherwise gone unnoticed, they make longitudinal qualitative studies problematic. The constant change was perceived as both a challenge and an opportunity by some of our interviewees: “The thing that I found to be one of the most challenging but [also] one of the most interesting things is that my site of study is constantly changing, which means things can be obsolete or completely meaningless by the time I get around to putting a paper together, but it is neat to watch, look at the process”.

The change in interface sometimes caused changes in the research context, and in the studied online cultures -- reflected in individual communications that were removed or altered, patterns of interaction that changed, users who appeared and disappeared, and links, data, and complete sites that vanished. Some tools or content are expected by default to last for only a relatively brief time (e.g. tweets), leaving the researcher to ponder how to collect and interpret these data. This necessitated researchers to develop strategies to overcome constant changes. Some decide to completely ignore the changes and focus on more general research questions (“What do you do with that? Do you ignore it? You ignore the actual changes and just talk to the people about their overall experiences”), others worked manually to recover some of the data that was removed or used web archives to reconstruct it, or built their own websites which they could control and tailor based on their specific needs. In any case, all of the researchers were deeply aware of the issues brought on by external changes and reflected continuously on the effect they have had on their research, even when they couldn’t find practical ways of overcoming them.

Applying ethical oversight. Large scale online environments presented ethical issues that should be considered in any research study, but because of scope and scale they are exacerbated. The features of large scale online environments, and especially social networking sites, caused many researchers to inadvertently collect interactions with users other than their “official” study participants. These data can be the result of chat and personal messages, comment threads, friends’ lists and tagged photos, among others, and present a cause for concern to the researchers: “Ethically it is questionable that we can do what we’re doing, basically I have participants that have not consented to be in my study because I am looking at other people’s profiles so I see their friends’ name and I see their status. They are participating without knowing, in a sense”.

All interviewees mentioned that current IRB standards do not sufficiently address the needs of researchers of large scale online environments and those of their participants. While being overly restrictive in some cases (“Our IRB wanted us to have every single member of a huge community opt [in]”), they didn’t offer proper guidance in cases of ethical conflicts. As one interviewee summed it up: “You have informed consent, that’s fine, but when the chat message comes up from somebody else and you capture it, you do not have their informed consent, you go back up to ethics school!”. However, no community-wide discussion of ethical issues pertaining to the specific characteristics of large scale online environments takes place nor are there golden ethical standards to guide researchers. Rather than reframing the discussion at a community level, each individual researcher or research group crafts their own provisional ethical standards, which range from adding a general message reflecting the presence of a researcher in the relevant online environment, to a mandatory no-use and no-documenting rule of information incidental to that obtained from the actual participants.
Theoretical issues

The second theme that came up from the interviews surfaced the core theoretical struggles that qualitative researchers of large scale online environments face. These reflections situated the work of qualitative researchers within epistemological and methodological choices. They could roughly be categorized as:

1. Valuing research questions over methods
2. Flexible methodological choices
3. Power structures within the broader research community

Valuing research questions over methods. Our interviewees defined themselves as qualitative researchers, but when asked about the research process they followed all advocated the primacy of the research question over the method: “What matters is the question and not the method. What matters is that the question gets answered. And the method that was used were [sic] the best we could do to get the most full answer within the parameters we set and the question we posed”. The reason for that was both pragmatic – the need to attain actionable results that will translate into an understanding of cognitive processes, but also an almost philosophical view of what constitutes effective research: “We are also solving real problems, and it is not useful to the users or to us if we try to solve a problem with the wrong tool, no matter how attractive the tool is.”

As predominantly qualitative researchers, they all favored exploratory studies led by relatively broadly scoped, open research questions, over pre-formulated hypothesis-driven research. The nature of large scale online environments lends itself well to open research questions that offer new paths for exploration. At the same time, such environments presented serious challenges to maintaining academic rigor: the possibility of delving deep into the multitude of interactions, and simultaneously getting a comprehensive, perspective, is seriously limited. To maintain and balance methodological rigor with the broad nature of exploratory research and the vast scope of these environments the researchers focused their questions on attainable goals. Some suggested building up the research questions gradually as the research progresses: “Practical is what it is. Think about the question. Ask it. How much time do you have? Well, how much data do you need to collect? Do that. Stop. Write. Ask again. Move to the next”.

Flexible methodological choices. The primacy of the research questions over the methods led to a surprisingly agnostic view of methods, and a liberal use of various methods that do not commonly typify qualitative research. Coming from qualitative backgrounds, all researchers shared the belief that “qualitative methods will yield more interesting data. It will yield a more ecological balance of results”. But in practice, as they were looking for answers to hard questions, implicit relationships and a sense of meaning in large scale online environments, the researchers complemented qualitative methods (e.g. interviews, observations, participant observation, artifact collection) with profoundly quantitative methods, ranging from purely statistical analyses, to content analysis and categorization, to natural language processing, machine learning algorithms, log analysis and social network analysis. The benefits of combining methods were obvious: “Some of us are very, very, tied to particular methods, but most of us are methodologically pretty flexible, because the big scale problems are going to need different ways of solving them”.

The acceptance of different methods stemmed from the researchers’ pragmatic view of their research being not only exploratory but actionable. Where the vast landscape of large scale online environments was the case, efficient ways of obtaining multiple points of view compensated for the diversion from purely qualitative methodological rigor. As one interviewee recalled: “I really think that the statistics by themselves are like breadth and we’re depth, and I think that the qualitative method always departs from this. A complementary use of qualitative and quantitative methods makes this study stronger”. Another said: “when you talk about convincing people about your methods, it is sometimes useful to have a little bit of quantitative data to sort of use it as a base for some of your arguments”.

This was also the outcome of an epistemological belief shared by many of the interviewees, where qualitative and quantitative methods offer different facets about the same phenomena: “I do not believe in a qualitative/quantitative distinction. Any quantitative thing is in relation [to] a qualitative set of assumptions or what constitutes a good analysis for this particular purpose. And so when people say to me: “ethnography always starts this way!”, in that you are a zealot”. Although various complementary methods were mentioned by the interviewees, first among them was social network analysis. Social network analysis was lauded for its ability to uncover discrete pockets of potential interest, but was not
accepted unconditionally - interviewees were quick to note the oversimplification of social network analysis products, especially with regard to uncovering cognitive processes, motivational factors and human values. They were troubled by reports of structure and ties that were not followed by an in depth analysis of the findings. One example of such an argument was - “I find there are a lot of social network analysis papers that come up with a profound conclusion: user 26 is most central in the network. And my response to that is: “and I care, why?””.

A social network analysis was viewed as a beneficial tool only in combination with deeper qualitative analysis, most interviewees looked in favor on forming multi-method research teams to tackle research about large scale online environments. They saw the benefit in such teams in the potential for mutual learning among team members coming from different research traditions, and in providing different data and understanding through multiple domain and methodological lenses, and deemed multi-method teams to be necessary to move research of large scale online environments to the next level:

“There is an acknowledgement of the need for additional points of view, but in general we all sort of agree that a successful project is going to have multiple minds, it is almost like seeing the argument that we made for participatory design back in the early 1990s, that if you have a hard problem you need multiple perspectives to solve it”.

Power structures. The combination of research question-oriented work and the relatively unorthodox methodological choices, were a point of contention and conflict between the researchers that were interviewed and some of their research communities.

These tensions stemmed from the juxtaposition of an ever-changing research field (both technologically and culturally), and the pragmatic research needs, on one hand, and the formalist structure of various methodologies, whether qualitative or quantitative, on the other. As was discussed previously, most, if not all, the researchers we interviewed chose to take a practical stance which placed an emphasis on the research question at hand; thus, they placed themselves in an intermediate position which does not fall into a schematic methodological tradition. As a result of that, the researchers faced push-back from both sides – to purely qualitative, and specifically traditional ethnographic researchers, they do not seem reflexive enough, while quantitative researchers frowned upon the lack of measurable data they offered.

The interviewees strongly defended their purposeful situation between the two research communities, as an appropriate epistemological choice when studying the ever-changing large scale online environments (“I do not pretend that we do the kind of ethnography that someone who goes to study the natives in Bali does. We do not have the time. If we could, that would be lovely. But the technology would change in the meantime”), and because of shifts in the patterns of interaction that result from the appropriation of the constantly-changing technologies and practices would also change. (“Anyone who comes to me and says that [we] do not need to know anything about the statistics and infrastructure, I’m like, great. You’re essentially a theorist. But what I’m looking for is a profound understanding of the shifting patterns of activity. If you’re interested in the shifting patterns of activity then you need to understand something about data usage and uptake”). All noted the various difficulties that they faced because of their placement between the different research communities.

Some of these difficulties were:

Lacking a sense of community within the broader research community – several interviewees discussed their sense of isolation when doing qualitative work on large scale online environments. Notions of internal struggles among researchers, lack of support for reflexive practices or, alternatively, use of multi-method approaches, contributed to this feeling (“I feel that in some way we’re afraid of our colleagues”).

Writing and reviewing process obstacles – issues related to writing styles, stylistic requirements, and the ensuing reviews, were a recurrent theme in all the interviews. Many researchers railed against the existing templates and stylistic requirements demanded by leading conferences as being overly-positivistic, and not amenable to qualitative/exploratory descriptions. These, by nature, are longer, and require more detail than empirical reports, as was noted by one interviewee: “I’m not particularly winning this battle, you end up with a massive amazing data, and telling people about one tenth of it through the kaleidoscope of this dreadful template and this ridiculous infrastructure”.

At the same time, the researchers criticized their very own qualitative research community for not crafting proper writing styles that would get the research contribution across to wider audiences: “Because we’re not writing it well, they’re not hearing it well.”
Limited dialog and marginalization of qualitative contributions – getting the message across to wider audiences proved to be even more difficult when researches sent their works for publication. Most of the dialog occurs through conference publication and the review process. Several of the interviewees lamented the lack of broader dialog outside of reviews, which they felt were tipped in favor of the empiricist/quantitative model, and rarely allowed for acceptance of exploratory qualitative works. As one interviewee succinctly put it: “If it is not positivistic it does not exist”. Getting the message across to reviewers proved to be a frustrating experience: “Reviewers don’t understand our methods and we get critical comments that are ill intoned. It is also just tedious to explain the method yet again. If you had someone doing quantitative methods who had to start explaining basic statistics in their paper they would tear their hair out. I feel like that is what we have to do.”

Yet again, the unique stance of methodologically agnostic researchers proved to be extremely difficult: “It is inherent in the interdisciplinarity of the field that you are going to meet people who fundamentally do not understand your methods and its [sic] basic assumptions.”

The lack of ongoing community-wide dialog on one hand, and the extremely critical dialog manifested through reviews, on the other hand, coupled with the relatively small number of researchers doing qualitative work in large scale online environments, led most of the interviewees to feel marginalized, in a way that made it necessary for them to continuously fight for the acceptance of their research and their practices.

Insights from the Meta-analysis of Published Papers

The meta-analysis we conducted supported some aspects of the findings from interview analysis. However, the themes brought up by our interviewees did not correspond to those that we found in the meta-analysis of the papers: most of the discussion of the qualitative methods that occurred in the published papers was limited to a less rigorous presentation (sometimes lacking details of the research design). The relationship between the research questions and the choice of methods was rarely discussed nor did we see much discussion of ethical issues, either in small or large scale online environments. There was very little discussion about how the researchers cope with challenges when doing this kind of work and how they felt. While it is interesting to reflect on these differences it is also important to remember that open-ended interviews encourage exploration of themes and feelings, while published papers tend to document how a method is used within the context of a specific study; this is especially true given the format and the length of published papers that, as our interviewees mentioned, did not lend themselves well to an in-depth discussion of qualitative work.

The meta-analysis was valuable for revealing a noticeable change in the research environments that were studied: while the use of ethnographically-inspired methods to study offline environments maintains its dominance over the years, online research gained popularity from 2002 onwards with mixed offline and online work maintaining an almost constant stream of 5-7 papers per year. From 2006 onwards, mobile environments became a popular area for ethnographically-inspired qualitative work, with about of 2-3 papers a year.

We were surprised to learn how few studies have used qualitative methods in large scale online environments. Within the 6 papers that focused on large scale online environments, the variation was wide in terms of the methods employed (from long and deep immersion in the field to brief observations), and so was the size of the group or culture studied (a specific environment (e.g., WoW) or cross cultural study). We believe that the findings from our interviews help in understanding this low number – the practical and theoretical challenges that researchers face when doing and writing qualitative work are exacerbated when trying to adhere to the publication practices common in the journals and conferences that were examined. It is possible that preferable venues for publication of such work are monographs and books, but these were not part of our analysis.
Conclusions

Our research has surfaced many challenges that come from the negotiation between the scope of the field, on one hand, and the focused nature of qualitative work, on the other. The practical issues the interviewees faced were the product of pairing various combinations of tried and tested methods that have been used in studying smaller scale settings, with large scale environments. These challenges affected the ways in which qualitative researchers of large scale online environments scoped the research setting, accessed both participants and data, navigated the ephemeral structure and culture of these environments, and maintained a proper level of ethical standards. The practical issues were set against a broader framework of theoretical challenges, speaking to methodological choices and power structures within the broader research community: for example, qualitative research of large scale online environments is not “pure” qualitative work. It stems from the research question and not from the methodological proficiency of the researcher; it is relatively methodologically agnostic and incorporates other, often quantitative, methods to allow for multiple perspectives and to complement the qualitative work; and, it suffers from power issues due to its standing between purely qualitative research and the positivistic tradition. While these challenges are by no means foreign to qualitative work, they are exacerbated by the scope and attributes specific to large scale online environments and they may be responsible for the low number of published papers on this topic.

Looking at these findings, we ask how can we as a community approach the challenges in a way that will support researchers doing qualitative work in large scale online environments and at the same time advance the broader research community.

One way is to address these issues by focusing on qualitative researchers of large scale online environments as a community of practice, and craft mechanisms to support them. Communities of practice are defined as “an informal aggregation of individuals engaged in common enterprise... characterized by the shared manner in which its members act and how they interpret events.” (Pawlowski et al., 2000). Indeed, from the themes that came up from the data, emerged a clear image of a community of practice that is important to the broader information science research community. The way qualitative methods are adopted, adapted, and practiced around large scale online environments demonstrates their plastic nature and their ability to carry meanings across communities sympathetic to this type of research. While this community of practice is a fluid one, as many researchers weave in and out as they go between different research projects of varying scale and type, it is also a very well defined community, with its own theoretical and practical foci, and a strong sense of isolation from the broader research community.
Focusing on the challenges outlined by our interviewees as having an impact not just on the immediate qualitative research community, but on the broader research community may also lead to a better understanding the relatively low number of papers discussing large scale online environments that we found in the meta-analysis. Unpacking these challenges and addressing them technically and theoretically is needed for qualitative research to flourish in these environments.

The challenges outlined by our interviewees also illustrate the need for a community-wide discussion through which we acknowledge that qualitative methods, along with quantitative methods needed to gain a deeper understanding of how people interact in large scale environments and how these environments evolve and change.

To do that, in a way that will be meaningful to the broader research community entails a growth process, which outwardly embraces qualitative methods as a valued and valid way of researching large scale online environments, and not just accept them as ancillary tools. This process of growth and development entails, by default, some growing pains which are bound to pass in time. Large scale online environments change patterns of interaction, and require new approaches and methods to research them. Formidable though this process may be, it paves the way to one of the new and most exciting areas for information studies research and presents a challenge that we can all embrace.

References


An Analysis of Cooking Queries: Implications for Supporting Leisure Cooking

Sally Jo Cunningham
Computer Science Department
University of Waikato
Hamilton, New Zealand
sallyjo@cs.waikato.ac.nz

David Bainbridge
Computer Science Department
University of Waikato
Hamilton, New Zealand
davidb@cs.waikato.ac.nz

Abstract
Cooking is a common and an information-intensive activity. We analyze a set of 678 cooking-related queries to identify the attributes that cooks provide in their queries to the Google Answers™ ‘ask an expert’ online reference system. The results suggest directions to take in developing an effective organization and improved functionality for a cooking-focused digital library.

Keywords: cooking, information needs, query analysis

Introduction
Cooking is a common, complex, information-intensive activity. The traditional information resource is the physical cookbook (delightfully referred to by Teng et al, 2011 as “the tombs of printed recipes … found in almost every kitchen”). Recipe and other cooking-related documents are widely available on the Web, and the support of cooking through improved access to cooking information is an active and growing area of research.

However, at present significant gaps exist in our understanding of information seeking behavior in regards to cooking: What are leisure (that is, non-professional) cooks looking for when they consult cookbooks, websites, cooking videos, etc.? How do they describe their information needs? What needs do they find difficult to fulfill? Answers to these questions may point to enhancements to current digital cooking resources—additional access points, functionality, and document and information types. This paper adds to our understanding of cooking information behavior through an analysis of a significantly sized (678) set of cooking questions posed to an online community reference system (Google Answers).

This paper is organized as follows: we first explore related work into cooking information needs and how they can be supported (RELATED WORK); we next describe the source, characteristics, and analysis method for authentic cooking information requests examined in this present paper (DATA GATHERING AND ANALYSIS); we explore the broad categories of information that the posters wanted (CHARACTERIZING THE DESIRED RESPONSE); we closely examine how posters describe one of the most commonly requested categories, the recipe, and suggest possibilities for enhancing recipe searching / browsing (DESCRIBING RECIPES); and we conclude by considering how a full-fledged cooking digital library could move beyond the bare-boned functionality of a recipe collection (SUPPORTING COOKING IN A DIGITAL LIBRARY).

Related Work
The work related to this study falls into three categories: ethnographic investigations of cooking information needs; examination of recipes and their interpretation; and examining usage of software that supports cooking.

Copyright is held by the authors.
Ethnographic Studies of Cooks and Cooking

A rich literature exists describing cooking related information behavior of serious but non-professional cooks (Hartel, 2006; Hartel, 2010a; Hartel, 2010b). These studies were conducted as small scale (20 to 31 participants) interviews and observations of self-identified gourmet / hobbyist cooks, with a focus on uncovering the time-based behaviors and processes in which cooking information behavior is embedded (Hartel, 2006; Hartel, 2010b), the management of personal cooking information collections (Hartel, 2010a), and the resources hobbyist and gourmet cooks consult (Hartel, 2006; Knopp, 2011). An earlier examination specifically of cooking information needs of hobbyist cooks (Knopp, 2011) relied on 24 descriptions of ‘cooking episodes’ elicited from study participants. In that study, the primary information requests were for recipes, explanations of cooking techniques, and where to obtain hard-to-locate ingredients. Through this previous research, a rich picture has been developed of the sources consulted by ‘serious leisure’ cooks (Stebbins 2007), the ways that they consume this gathered information, the points in their cooking-related activities at which they seek further information, and how they build and maintain their personal information collections.

Less is known about information behavior and needs of novice cooks or to support pedestrian, ‘everyday’ home cooking. Van de Wiel et al (2010) explore the attitudes to cooking, cooking habits, and cooking / eating contexts of young male professionals with low levels of cooking skills, through contextual interviews of 11 participants. Key findings motivated feature development to the CookKing system, including supporting social and competitive aspects of dish preparation and adding mobile functionality to aid in ingredient shopping. A diary study of five young adults with limited cooking experience (Palay & Newman, 2009) elicited similar results, and also identified the need for assistance in achieving greater variety in meals.

These studies adopt intensive qualitative methodologies (diary study, interviews, observations), and so necessarily are based on limited numbers of participants. This present paper examines a significantly larger data set (678 cooking related queries posted to the Google Answers forum)—but far less background and contextual information is available for these queries than with the earlier studies. For example, the Google Answers queries differ from the information needs as described in (Knopp, 2011) in several significant ways: the postings are not necessarily tied to a specific cooking event (that is, a query could be part of an ongoing information gathering exercise and was not necessarily tied to creating a particular dish or meal), and the Google Answers posters themselves had not been able to satisfy their information need (whereas the participants in (Knopp, 2011) were asked to describe information needs for a successful cooking event, and so presumably their questions had been answered).

These earlier studies targeted participants with a common cooking skill level (eg, gourmet hobbyists, young men with limited cooking experience). By contrast, the queries from this present study span the spectrum from confident gourmet hobbyists (eg, “I’d like to make high-quality sushi at home”), to the cautious experimenters (eg, “How do you make Hamburger Helper-Cheeseburger Macaroni from scratch?”), to utter novices:

I’m a 19 year old guy living in Singapore with nothing to do for 2 weeks. I am thinking of baking/cooking for fun! What should i know/have before i start on my adventure? What are the tools involved? How do i bake a cookie?

Examining Recipes and Their Interpretation

Another approach to supporting information behaviour in cooking begins with a close examination of information presentation structures in recipes, and how people understand (or misunderstand) them. Hamada et al (2000) and Shirai & Ookawa (2006) analyse collections of Japanese recipes to develop models of the steps involved in cooking, with the goal of automated processing existing recipes to produce multimedia / animated ‘walkthroughs’ of how to cook a given recipe. Unacceptably high rates of errors in automated extraction of cooking actions arise from variation in vocabulary across recipes—highlighting the difficulties people face when learning to cook. One aspect of this vocabulary problem is examined by Yamakata et al (2009): how people naturally refer to ingredients in intermediate stages of preparation (eg, “minced onion”). Here a surprising variation in vocabulary of the participants was identified, compounding the variation in recipe vocabulary noted by the earlier research. Though this was a small-scale study (involving 20 participants), its results are promising in suggesting query processing techniques to match a searcher’s native vocabulary to existing recipes, over all steps in a recipe.
Van Pinxterin et al (2011) investigate how people perceive similarity in recipes, through a card sorting study (14 participants, sorting 66 recipes). A model was induced of the most significant characteristics contributing to perceived similarity, and this model was incorporated into a cooking support system that recommended ‘healthy’ dishes similar to earlier recipe selections of the participants. The authors note that their recipe model was ‘quite different from other approaches in the literature’.

These studies attempt to leverage off the large number of existing, digitized recipes, to identify techniques to make these recipes more useful, usable, or accessible (particularly to novice cooks). Further research on human perceptions of recipe structure and vocabulary is needed to more effectively support recipe searching and the searcher’s understanding of recipe search results; matching between the ‘pain points’ identified through qualitative explorations of information needs may be useful in guiding this research. However, this exclusive focus on recipes misses out on the prospect of supporting behaviors associated with other significant cooking related documents used by cooking hobbyists (eg, culinary magazines, cooking videos, menus).

Learning from Deployed Cooking Information Systems

A third approach to investigating cooking information behaviour is to examine how existing cooking document collections are used, for example through usage log analysis. Despite the long-standing and widespread availability of large, digital recipe collections, no studies have been conducted to date of how people search and browse these collections—a serious gap (and surprising) gap in the research literature.

However, evaluations of novel features implemented in prototype cooking information systems provide insight into promising directions for further investigation. Users of the CookKing system (van de Wiel et al, 2010) found their individual histories of favorite recipes to be useful in meal planning and developing shopping lists. A small-scale case study indicates that an individual’s recipe browsing and cooking history can also serve as the basis for a model of their preferences, to inform personalized recipe recommendations (Ueda et al, 2011).

Evidence of other users’ preferences, opinions, and recommendations can be useful when searching and browsing for recipes. A six month study of the experimental Kalas social navigation system indicates that users both liked these “social trails” through a recipe collection, and also modified their searching/browsing through Kalas to take advantage of them (Svensson et al, 2005).

Data Gathering and Analysis

The source of the natural language cooking-related queries analyzed in this paper is Google Answers’ ‘ask an expert’ service (http://answers.google.com). In this service past queries and their answers are publicly available, arranged into categories (with the person who posts the query selecting the appropriate category for it). This is an attractive source of ‘real’ questions, and allows us to study active information needs rather than post-hoc recollections (as in Knopp, 2011).

The questions posted to Google Answers are in natural language, describing the information need rather than being restricted to a few user-selected query terms (as is the case in most digital libraries and search engines). For example, an analysis of approximately 15 million MSN search queries found that 90% of queries include four terms or fewer, and 99.9% of queries include twelve terms or fewer (Bendersky and Croft, 2009). In contrast, the Google Answers queries average 71 words. The Google Answers user places a monetary value on the question (the bounty offered was limited to be between US $2–200) and pays the expert on receipt of a satisfactory answer; these are thus significant rather than casual questions. The complete set of the Food and Cooking category to Google Answers’ ‘ask an expert’ system (http://answers.google.com) was downloaded: 678 queries posted to the site from 10 April 2002 to 23 November 2006 (when the service was discontinued). The questions posted to Google Answers are in natural language, describing the information need rather than being restricted to a few user-selected query terms. Queries averaged 71 words.

The complete set of the Food and Cooking category to Google Answers was downloaded; the set totaled 678 queries posted to the site in the period 10 April 2002 to 23 November 2006. The value of postings ranged over the full limit the site allowed with a mean price of US $11.62 (standard deviation of 18.3), and mode and median value being the same, at US $5. Of the 678 questions posted, 49% of them
were successfully answered, meaning the originator of the question accepted the answer, and paid the fee.

Figures 1, 2, and 3 provide a more detailed overview of the pattern to the posed questions and answers, and are useful in determining clusters and extremities in the data.

**Figure 1.** Query length in words versus time taken (in minutes) to answer the query.

**Figure 2.** Rating of answer versus Query Length (in words).
In Figure 1 we get a sense of how long a user needs to wait to get an answer. Almost half (49%) are answered within the first two hours, rising to 63% answered within the first 4 hours, and 85% within 24 hours. Through this 2-4 hour band we see higher activity in the y-axis (length of query), showing that taking the time to provide a posting with some detail pays dividends in terms of getting a timely answer. The quickest response time was one minute for a question asking about the shelf life of eggs in a refrigerator.

Google Answers posters are given the opportunity to rate the quality of the answers provided. As Figure 2 shows, most users are highly satisfied with the answers given, rating the answer either with the highest or second highest rating. Some lower ratings were assigned to answers where the questions asked themselves were rather brief (under 100 words) compared to others.

Turning our attention to the answers given, Figure 3 visualizes the average number of words in the response to a query against the speed of response. The norm for answers is under 2000 words, with a rising tail indicating that longer answers (not surprisingly) tend to take more time to compile. The longest query was over 14,000 words long. This was a response to a question asking for a list of the top 1000 recipe books based on sales. From this phase of the analysis, 93 of the 678 queries were discarded as being off-topic to this present study of hobbyist cooks: for example, queries about cutlery, china, and glassware ("Where can I can definitely buy this whiskey / whisky glass? http://tinyurl.com/7kmna"), queries about commercial equipment ("I am looking for information about how commercial soft serve ice cream machines function, in technical terms."), or advice on starting a food-related business ("Mum wants to sell a little of her home-made cakes, to local shops, in Australia?").

A grounded theory approach (Glaser & Strauss, 1967) was taken to analyze the remaining 586 queries. The queries were coded to characterize the desired response to their query (eg, a recipe, an explanation of a cooking technique). Recipe-oriented queries were further coded to characterize the details that users provided when describing the desired recipe (existing recipe ontologies (Dale, 1989) and (Kimura et al, 2008) were consulted to suggest labels for emergent categories). The coding scheme is described in the RESULTS section.
Characterizing the Desired Response

As summarized in Table 1, the Google Answers posters were seeking a broad range of responses:

Table 1  
Breakdown of Desired Response Types

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product Information</td>
<td>213</td>
<td>36.41%</td>
</tr>
<tr>
<td>Recipes</td>
<td>119</td>
<td>20.34%</td>
</tr>
<tr>
<td>Hygiene, Health, Food Storage &amp; Preparation Advice</td>
<td>53</td>
<td>9.06%</td>
</tr>
<tr>
<td>Food Trivia and Market Research</td>
<td>42</td>
<td>7.18%</td>
</tr>
<tr>
<td>Definitions</td>
<td>39</td>
<td>6.67%</td>
</tr>
<tr>
<td>Food History and Background</td>
<td>37</td>
<td>6.32%</td>
</tr>
<tr>
<td>Resource</td>
<td>34</td>
<td>5.81%</td>
</tr>
<tr>
<td>Troubleshooting</td>
<td>27</td>
<td>4.62%</td>
</tr>
<tr>
<td>Menu and Food Combination Suggestions</td>
<td>23</td>
<td>3.93%</td>
</tr>
<tr>
<td>Cooking Techniques</td>
<td>21</td>
<td>3.59%</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>20</td>
<td>3.42%</td>
</tr>
<tr>
<td>Measures, Conversions, and Substitutions</td>
<td>15</td>
<td>2.56%</td>
</tr>
<tr>
<td>Contacts for Experts</td>
<td>5</td>
<td>0.85%</td>
</tr>
</tbody>
</table>

Note: Does not sum to 100%; some queries include more than one desired response category.

**Product information.** Postings request information about a food item or physical equipment involved in cooking.

Food items ran the gamut from the 'gourmet' ("Les Langues Dorees’ Demel Vienne chocolates") to the humble ("frozen limeade concentrate"). These edible items might be ready-to eat ("Nutria jerky"), an ingredient the poster anticipates to use in his/her own cooking ("sausage skins for home-made sausages"), or a sauce or other accompaniment to a dish ("Looking for how to get a salad dressing named chateau classique. Maybe distributed by Sexton Foods, Inc.").

With food items, posters wished to purchase products they were finding difficult to source—for example, a treat from one’s country of origin ("Where can I buy LUCKY CHARMS breakfast cereal in the UK?"); a specialist version of a common product ("Looking for fresh farm lamb and other quality meats such as farm pork, farm free range chickens with no growth hormones and antibiotics"), or a commercial product for which no brand name was known ("How can I obtain Powdered alcohol in the united states?"). Less frequently (four queries), if the item itself could not be located then posters requested a substitute ("I had a liquor in Curacao named Ponche Kuba…If it isn't available what other liquor can be substituted for it instead?").

The majority of equipment related posts were seeking to purchase the item. The types of equipment desired included replacement parts, specialist food service and prep gadgets, ovens, and food storage devices. In contrast with the food item requests, fewer of the equipment requests included a brand name of the desired item—the poster instead provided a more-or-less elaborate description of the desired item. The descriptions ranged from the straightforward request for a generic instance of an item ("Where can i buy a JELLY MOLD in toronto?") to detailed requirements:

I am looking for a recommendation on a blender to bring to work to mix up shakes. Being that this is a typical office environment I need something that won’t disturb my coworkers.

Requirements:
1) Super Quiet
2) It is for making shakes with ice and water. (needs to be able to crush ice)
3) Would be nice if it were the one piece types that are easier to clean. (not entirely one piece, just the top glass portion)
4) Looks are unimportant
5) Order online for less than $100

Clarification: I’m looking for a standard blender, not a hand-help model.
Of the equipment queries not requesting purchase information, three queries related to safety or hygiene issues with cooking gear (e.g., “is it wrong to run microwave ovens with nothing in them and if yes, why”), and three requested plans or advice in constructing a specialist cooking device (a brick oven, an elaborate wooden outdoor pig roaster, and an apartment-based sake brewing environment).

**Recipes.** Recipes are requested for complete dishes (e.g., “green chilli mashed potatoes”), a sauce or other accompaniment to a meal (e.g., “Japanese curry sauce”), a food item that can be used as an ingredient in other recipes (e.g., “Mozzarella cheese”), and for common sub-parts of more complex recipes (“A good old-fashioned pie pastry recipe required, savoury/meat etc, not sweet!!”). Recipe queries can further be loosely divided into requests for ‘a’ recipe (that is, any recipe meeting the conditions of the query, “a recipe for Halva”), and requests for ‘the standard’ or ‘the best’ recipe for a specific dish. How posters described recipe requests is examined in the next section, DESCRIBING RECIPES.

**Hygiene, Health, and Food Storage & Preparation Advice.** These queries include requests for good practice for food handling (“I am a cheapskate, with a lot of uncooked meat which is probably [sic] not ‘bad’ yet. But I am not certain. If I freeze this meat, will it kill any ‘bugs’ or bacteria already in the meat? Or, am I simply delaying my trip to the Tomaine Temple?”) and background on potential health effects of food preparation techniques (“Does microwaving food in plastic containers or plastic cling wrap release harmful chemicals into the food?”).

**Food Trivia and Market Research.** These queries were directed most frequently at ‘top ten’ type lists (“I am looking for the top ten countries in the world in terms of the consumption of clams”) and statistics on purchase and dining trends (“where can i find statistics on whether people leave or drink the milk after finishing all the cereal in the bowl?”; “there are many different kinds of miso in the stores. How do I know which is most like the miso commonly served in Japanese restaurants in North America?”).

**Definitions.** Posters request: an ‘official’ name for a specific food item or technique (“A raw egg swallowed in Sherry and Worcester Sauce: how is this hangover cure better known?”); a description or definition of a food item (“I was in the indian grocer this afternoon, and a bunch of people were lined up waiting for ‘tawa’. What is “tawa”?"), and an explanation of the difference between two related food items (“What are the differences between Caramel and Butterscotch?”).

**Food History and Background.** These posts include general requests for ‘history’ of a food item or technique (“What is the history of teppanyaki cooking?”), as well as more specific inquiries about origins and practices (“Afghan Biscuits are a popular biscuit/cookie in New Zealand. How did they get their name?”).

**Resources.** Posters requested print or digital resources that they could use to satisfy further personal information needs: for example, cookbooks, recipe websites, discussion forums, and more specialist cooking and food related websites (e.g., “Strawberry themed cake images”; “a webpage that lists all produce that supermarkets sell, complete with pictures”; “a website that will design a menu based on a few budget friendly ingredients so that you get the most bang for your buck.”).

**Troubleshooting.** Posters requested information on: how to recover from a cooking disaster (“What takes the burnt flavor from rice when you burn it?”); the underlying causes for a cooking problem experienced in the past and how to avoid them in future (“What mistake did I make when I baked challah bread that was too dense? I followed the recipe perfectly and removed the bread a few minutes before the allotted time.”); and tips for achieving more professional cooking results at home (“How do you cook green vegetables and have them stay green? My green beans always get pale, yet I have had vivid green beans in restaurants.”).

**Menu and Food Combination Suggestions.** These posts request suggestions on appropriate items to create or complete a specific menu (“I want to make a list of standard spice combinations, correlated with the types of dishes the spice combination works with”; “What foods go with Chianti Wine”), or request more general advice on food and ingredient combinations. (“Do things such as honey, coconut
and the like go well with chicken? What other sweet things can be used to compliment a good piece of chicken?"

**Cooking Techniques.** These posts request an explanation or description of a cooking technique or food preparation, rather than a recipe per se ("I'm trying to discover the secret to getting steaks to come out the way they do in restaurants if you order them medium rare, with a smooth dark brown crust on the outside but medium rare (light pink throughout the inside."). Most technique references are to everyday home cooking situations ("What is the best way to boil a perfect hard-boiled egg?"; "I would like suggestions on how to peel a butternut squash without becoming an amputee.")., but some are more specialized and idiosyncratic ("In an Episode of Father Ted Mrs Doyle gives a cake to Eoin MacLove with a sweatshirt inside it. How do I make a similar cake with a sweatshirt baked in the middle of it?").

**Miscellaneous.** Queries that do not fit into any other category ("I am thinking of baking/cooking for fun! What should I know/have before I start on my adventure?").

**Measures, Conversions, and Substitutions.** Conversion queries include present day measures ("What's is the approximate volume of 250 g wheat (or rye) flour?") as well as antiquated measures ("I have a recipe for Devil’s Food Cake from my great grandmother that calls for ‘1/3 cake chocolate’, and I want to know how much powdered cocoa or unsweetened baker’s squares to substitute."). This category also includes queries relating to the calories or dietary units in food items and substitution of food items (eg, Weight Watchers diet ‘points’) and substitution of food items ("Can red and yellow saffron be used interchangeably?").

**Contacts for Experts.** These queries are satisfied by names and / or contact details for human experts in various aspects of cooking: food critics, chefs, and wine experts.

**Describing Recipes**

In its minimal form, a recipe consists of a title, a list of ingredients, the amount of each ingredient, and a set of instructions for processing the ingredients to prepare the named dish. Here, we analyze the descriptive factors posters use to characterize recipe requests (Table 2).

In many print cookbooks and digital recipe repositories, the most common access points (searching and browsing) to the collection are the name of the dish, the main ingredient, and the course or meal associated with the dish. These three are the most frequently occurring descriptor types in the Google Answers queries. However, specifying these might not be straightforward.

There might be no ‘official’ name for a dish, or no single standard English spelling for a dish originating in another culture. A dish name may reflect its appearance (eg, Toad in the Hole, an English standard involving sausages in batter) or folktale / cultural associations (The Imam Fainted, a Turkish eggplant dish) rather than its ingredients. These difficulties are common enough to suggest that query interfaces should not rely overly heavily on recipe name searching. While prototype systems exist that try to work around difficulties in named dish searching by, for example, matching to query descriptions of a recipe’s process structure (Wang et al, 2008), these systems are far from deployment.

On the other hand, abandoning recipe title in favor of simple keyword search over the entire recipe text (as is common in current online recipe collections) can be frustrating as well, when the user wants a recipe for a food item that is commonly an ingredient in other dishes (eg, mozzarella cheese, sausage). In these cases, the option of searching by name / title can be invaluable. The usual difficulties with ‘bibliographic’ query pertain as well—users’ spelling may include errors, and they may misremember the name or other significant metadata for a recipe (eg, "Spelled to sound like “Padiusca” for puttanesca"). Spell-checking and browse-able term lists would be useful in addressing these problems.
Table 2
Characteristics Used to Describe Recipes in Queries

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
<th>Count(%)</th>
<th>Category</th>
<th>Example</th>
<th>Count(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of dish/item</td>
<td>Raisin tea cake</td>
<td>67 (56%)</td>
<td>Dietary restrictions</td>
<td>Vegetarian, gluten free, no dairy</td>
<td>15 (13%)</td>
</tr>
<tr>
<td>Ingredient</td>
<td>Bean sprouts, meat, roasted chicken</td>
<td>64 (54%)</td>
<td>Cooking technique</td>
<td>Dry curing, barbecue</td>
<td>15 (13%)</td>
</tr>
<tr>
<td>Type of dish, course, meal</td>
<td>Pasta, appetizer, breakfast</td>
<td>34 (29%)</td>
<td>Similarity to an example</td>
<td>Similar to a Mounds bar</td>
<td>15 (13%)</td>
</tr>
<tr>
<td>Epoch</td>
<td>Japanese, UK, Montreal</td>
<td>29 (24%)</td>
<td>Difficulty level</td>
<td>Easy</td>
<td>12 (10%)</td>
</tr>
<tr>
<td>Commercial, restaurant, chef</td>
<td>Coco Ichiban Curry House, George Shearing</td>
<td>29 (24%)</td>
<td>Print/Web resource where 1st located</td>
<td>Betty Crocker Creative Recipes: 1995-1997</td>
<td>10 (8%)</td>
</tr>
<tr>
<td>Time period or date</td>
<td>1950s, a year ago, ancient</td>
<td>22 (18%)</td>
<td>Holiday, event, special occasion</td>
<td>Valentine’s Day, tailgate party</td>
<td>6 (5%)</td>
</tr>
<tr>
<td>Sensory aspects</td>
<td>Crispy, buttery, flaky</td>
<td>21 (18%)</td>
<td>Procedure</td>
<td>Cut it up and boil it</td>
<td>5 (4%)</td>
</tr>
<tr>
<td>Appearance</td>
<td>Yellow, round, tapered</td>
<td>16 (13%)</td>
<td>Cost</td>
<td>Cheap, cost-effective</td>
<td>3 (3%)</td>
</tr>
<tr>
<td>Locally available ingredients</td>
<td>Available in UK</td>
<td>16 (13%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Does not sum to 100%; most queries include more than one information need descriptor.

Ingredients can be tricky to specify: can I say ‘meat’ or must I enumerate all locally available possibilities (‘beef or chicken or lamb or pork’)? Does the recipe use American terminology or English (eg, ‘bell pepper’ vs. ‘capsicum’)? Prototype systems presented in (Kimura et al, 2008) and (Liu et al, 2005) demonstrate the feasibility of ontology support for queries including ingredient specification.

Queries seeking recommendations for suitable dishes to serve frequently specify a course or meal, a holiday/event/special occasion, ingredients that the cook wishes to use (or use up), or degree of similarity to another dish that the person has tried. Recipe recommender systems are an active area of research (eg, Teng et al, 2011; Ueda et al, 2011; van Pinxteren et al, 2011) and can incorporate these facets of a need description.

A quarter of queries are for a specific version of a dish: the version created or popularized by a commercial organization, restaurant, or chef. These queries typically state that the user will not be satisfied with any other versions of the dish—suggesting that the ‘author’ or originating organization be included as metadata to recipes when this information is available.

Nearly a quarter of queries specify an ethnic, national, or regional origin for the desired recipe. Of course, all dishes come from some cultural context, so the remaining three-quarters of recipes are assumed to pertain to the dominant nationality of the Google Answers users (USA). The cultural assumptions for the target audience of a recipe (as distinct from the origin of a recipe—eg, a Korean recipe written for an American audience) should be made clear, as the measurements, names for ingredients, and common cooking techniques can vary across cultures (Kimura et al, 2008).

Similarly, posters requesting a recipe whose origins did not match the poster’s current physical location recognized that they might run into trouble sourcing ingredients, and so specified that they wanted recipes only with locally available ingredients. It is not too far-fetched to posit a system that could match a user’s current location to the inventory list of local online grocery stores, and use that information to remove or down-rank candidate recipes whose ingredients would be difficult to source.

Time period appears most useful when the poster is seeking historic recipes (eg, “ancient and traditional methods for curing meat”) or when tied to an iconic menu of a well-known restaurant or chef. Precise dates would be overkill—these queries are better served by broad categories such as ‘Middle Ages’ or ‘1950s’.

Sensorial aspects of desired dish (including texture, ‘mouth feel’, and intensity of flavor) address the impact of a dish on the diner—and so may be particularly useful in selecting between alternative recipes for a given dish. The system described in (Liu et al, 2011) supports search refinement by sensorial aspects by reasoning from a large-scale database of sensorial facts about ingredients.

The appearance of the finished dish is most frequently (and most elaborately) described when the poster hopes that something distinctive in the desired dish’s form will assist a respondent in identifying it, or allow the respondent to distinguish between similar dishes. Providing accompanying photos of finished dishes to recipes could be useful for these posters, as an aid in browsing potentially relevant recipes.
Current recipe systems put the onus on the creator or recorder of the recipe to indicate whether a specific recipe meets special dietary restrictions. Given that many such restrictions are defined by the elimination of a category or type of ingredient from recipes (e.g., vegetarian, no dairy), it is not far-fetched to imagine a system that can automatically infer whether a recipe meets the standards for common diet regimes—and indeed, the system described in (Liu et al., 2005) prototypes this facility. Less common restrictions may be supportable by negations in queries (e.g., one poster’s son found cream cheese to be indigestible—‘not cream cheese’).

Queries may include the cooking techniques and procedures for a dish when the poster does not know the dish’s name. Search over these facets can be supported by ‘mining’ the instructions for a recipe to create a representation of its workflow, and then comparing that to the query workflow (Wang et al., 2008).

When difficulty level is specified, posters ask for a simple, easy to prepare recipe that requires no specialist equipment—suggesting a binary rating of difficulty rather than more elaborate schemes. Similarly, the few queries specifying costs uniformly ask for cheapness, rather than finer-grained cost ratings.

Posters include surprisingly precise details of the print or web resource in which a ‘lost’ recipe was originally located. In the case of print resources, the difficulty is generally that the document has gone out of print—a problem that likely will be eased as more documents are digitized (possibly as individuals violate copyright to post favorite recipes). Links to Web recipes may break as the websites disappear or the site re-organizes, pointing to a need for personal digital collection support (Hartel, 2010a).

Supporting Cooking with a Digital Library

The above analysis clearly shows the needs of the Google Answers posters are wide-ranging. In terms of how a digital library infrastructure could help support their cooking information needs, our analysis suggests that the identified activities map well to a managed set of digital library collections, where cross collection searching and browsing the “super-structure” that matches the categories developed through ground truth analysis is supported. In addition the digital library environment should be augmented with embedded analytical tools, and support social structures such as forums and blogging.

For instance, the Food History and Background, and Recipes collections could be populated by the aggregation of existing on-line resources dedicated to cooking. If a protocol such as OAI-PM (Buchanan et al., 2005) were to be used, then these distributed resources could be regularly scanned for new and corrected content, and updated accordingly. Such co-operation of the primary source providers seems likely if—having located an item through the aggregated central collection—the user is taken to the original site so due recognition (and any relevant revenue stream) is given. If such an initiative were supported by Amazon (with their "look inside" capability) the Resources collection could be developed along similar lines, helping bridge the gap between the physical and digital forms hobbyist cooks are known to draw upon. Similarly a Product Information collection could gain significant leverage from integration with shopping aggregator sites such as Google Products.

The Cooking Techniques collection would work well as a video collection with content sourced from sites such as YouTube or commercial content providers. The Definitions collection could be seeded by text mining for facts in the other collections in the digital library. Tools embedded in the digital library environment would be essential, none more so than a measurement conversion capability. Coping with domain specific esoteric or dated forms of measurement that come up from time to time would be an interesting challenge—perhaps another opportunity for text mining to assist, this time by being applied to the Food History and Background collection.

Supporting Menu and Food Combination Suggestions is perhaps the biggest reach in terms of what is available in off-the-shelf digital library software. However, existing recipe recommender systems (e.g., Teng et al., 2011) could be integrated into a recipe digital library. Forum(s) for discussion would fill out the user experience. Our analysis suggests there is a niche for a specialist Trouble Shooting area where people can post and expect a speedy response.

We also noted earlier that hobbyist cooks also maintain a personal collection of resources, and that usage analysis of the CookKing (van de Wiel et al., 2010) and Kalas (Svensson et al., 2005) prototype systems indicate that users value personal histories of favorite recipes. A final “step-up” to what the digital
library provides, then, is to provide registered users with a personal space within the system. Given the approach details, this would be most naturally accomplished as a personal digital library collection. Using the idea of Digital Library Talkback (Bainbridge & Witten, 2011) a two channel can be formed: documents located by a user in the main (publicly available) collections can be transferred to their private collection; and counterpoised to this, content developed within a user’s private collection can be exported (published) and included in the publicly available collections.

Conclusion

By analyzing authentic queries to Google Answers, we have elicited a rich framework suggesting the broad range of responses that people want when they search for cooking information, and the types of metadata that they are able to provide when searching specifically for recipes. These posts represent user needs unfiltered by expectations of how a retrieval system expects a query to be presented, the access features that it supports, or the documents it contains. This helps to establish a clearer picture of what users really want and need in a cooking information system.

References


Empirical Analysis of Data Breach Litigation

Sasha Romanosky  
Heinz College  
Carnegie Mellon University  
sromanos@cmu.edu

David Hoffman  
Beasley School of Law  
Temple University  
hoffmand@temple.edu

Alessandro Acquisti  
Heinz College  
Carnegie Mellon University  
acquisti@andrew.cmu.edu

Abstract

The surge in popularity of social media, cloud computing, and mobile services has created an unprecedented opportunity for the collection, use and sale of personal consumer information. While these services provide many benefits, individuals suffer when this information is lost, stolen, or improperly accessed, causing them to seek legal redress. However, very little is known about the drivers, mechanics, and outcomes of these lawsuits, making it difficult to assess the effectiveness of litigation at balancing organizations’ usage of personal data with individual privacy rights. Using a unique database of manually collected lawsuits, we analyze court dockets for over 230 federal data breach lawsuits from 2000 to 2010. We investigate two research questions: Which data breaches are being litigated? Which data breach lawsuits are settling? By providing the first comprehensive empirical analysis of data breach litigation, our findings offer insights in the debate over privacy litigation versus privacy regulation.

Keywords: data breach, identity theft, privacy, litigation

Introduction

The surge in popularity of social media, cloud computing, and mobile services has created an unprecedented opportunity for the collection, use and sale of personal consumer information. While these services clearly provide many benefits to producers and consumers, individuals suffer harm when their personal information is lost, stolen, or improperly accessed, causing emotional distress or monetary damage from fraud and identity theft.\(^1\) Since 2005, an estimated 543 million records have been lost from over 2,800 data breaches,\(^2\) and identity theft caused $13.3 billion in consumer financial loss in 2010 (Bureau of Justice, 2011). In response, federal legislators have introduced numerous bills that define appropriate business practices regarding the collection and protection of consumer information,\(^3\) and federal regulators have drafted privacy frameworks for consumer data protection (Department of Commerce, 2010; Federal Trade Commission, 2010). For instance, the Department of Commerce inquired: “should baseline commercial data privacy legislation include a private right of action?” (Department of Commerce, 2010, 30). At issue is the degree to which federal consumer litigation deters privacy harms, or whether a new federal privacy statute is required.

---

\(^{1}\) See Solove (2007) for a description of the potential harms associated with breaches of personal information.


\(^{3}\) For example, the Cyber Security and American Cyber Competitiveness Act of 2011 (S.21), the Data Security and Breach Notification Act of 2011 (S.1207), the Commercial Privacy Bill of Rights Act 2011 (S.799), the Personal Data Privacy and Security Act of 2011 (S.1151), the Data Breach Notification Act (S.1408), the Personal Data Protection and Breach Accountability Act of 2011 (S.1535), the Secure and Fortify Electronic Data Act of 2011 (H.R.2577), the Cybersecurity Enhancement Act of 2011 (H.R. 2096).

Acknowledgements: This research was supported by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office, by Temple Law School’s Conwell Corps Program, and by the Information Law Institute at New York University School of Law. We would like to thank Antima Chakraborty, Carol Anne Donohoe, Ian Everhart, Caitlin Jones, Kevin Leary and Jake Oresick for their research assistance. We would also like to thank Paul Bond, Aaron Bumstein, Fainna Kagan, Amelia Haviland, Mark Melodia, Kristen Matthews, Peter Oh, Barrie Nault, David Navetta, Mohammad Rahman, Theresa Romanosky, Boris Segalits, Brendon Tavelli, and 7 anonymous attorneys for their valuable insights and suggestions.


Copyright is held by the authors.
However, little is known about the trends in data breach litigation – which breaches are litigated and which are not, and with what outcomes. Current scholarship examines only a narrow subset of lawsuits, usually focusing on high-profile cases or those with published opinions. And so, to our knowledge, no empirical research involving data breach lawsuits has been conducted. The purpose of this manuscript is to explore two questions. First, what kinds of data breaches are being litigated in federal court, and why? Second, what kinds of data breach lawsuits are settling, and why?

Overall, we believe this research can be of use to various parties. First, it can help provide firms with prescriptive guidance regarding the relative chances of being sued, and having to settle. This research could also be useful to insurance markets as a means for assessing and pricing cyber-insurance policies. Moreover, we believe that this work can help inform both plaintiff and defense attorneys in better understanding overall trends of data breach litigation. Finally, we hope that our research can inform the policy debate and help create a balanced privacy framework protecting both the interests of consumers who provide personal information, and organizations that collect and innovate using this information.

Related Work

In recent years, economists have researched a number of empirical and theoretical aspects of data breaches, such as the effect of breaches on a firm’s stock market price (Campbell et al., 2003; Cavusoglu et al., 2004; Acquisti et al., 2006; Kannan et al., 2007; Gordon et al., 2011), the effect of data breach disclosure laws on identity theft (Romanosky et al., 2011), and the conditions under which disclosure laws may reduce the social costs of these breaches (Romanosky et al., 2010). An emerging body of legal scholarship also analyzes court dockets. This form of empirical research makes very practical use of publicly available -- and generally very detailed -- collection of pleadings, motions, rulings and administrative record keeping that compose a legal dispute (Kim et al., 2009; Hoffman et al., 2007; and Boyd and Hoffman, ND). Intuitively, economic analysis of litigation suggests that individuals are more likely to file suit when their expected rewards exceed their expected costs (Cooter & Ulen, 2008, 414-484; Cooter and Rubinfeld, 1989). This hypothesis has been supported by some empirical work (Clermont and Eisenberg, 2002), especially in the area of financial patent litigation (Lerner, 2010).

Data

This manuscript combines a number of datasets described below. For the purpose of this manuscript, a data breach is defined broadly as the unauthorized disclosure of personal information by an organization.

Data Collection

To address our first research question ("Which breaches are being litigated?"), we first gathered a list of reported US data breaches from the Open Security Foundation ("Datalossdb"), a non-profit organization devoted to collecting and recording data breaches and IT vulnerabilities. Then, we used Westlaw to identify which of these reported breaches resulted in federal litigation.

To address our second research question ("Which data breach lawsuits settle?"), we used Westlaw to perform a systematic search for all federal lawsuits in which plaintiffs alleged an unauthorized disclosure of their personal information. (The lawsuit observations previously used are, of course, a subset of the results from this search.) Specifically, we searched Westlaw's Pleadings database using the following search strings: "personally identifiable information," "personal information," and either "data breach," "security breach," or "privacy breach." These search terms balance specificity without biasing search results to specific causes or types of data breach lawsuits. We then manually examined the results

---

4 These data are used per the OSF license agreement which states: “permission is granted to use this database in non-profit works and research.”
and extracted those cases relating to unauthorized disclosure of personal information. We believe this is
an appropriate combination of methods for identifying all lawsuits either filed in, or removed to, federal
court and therefore represents the most complete collection of federal data breach lawsuits.

We then used PACER to retrieve the court docket for each case. From the docket itself we coded
the following information: presiding judge, date filed, date terminated, forum, the law firms involved in the
suit and number of docket filings. We then purchased the complaint (or amended complaint where
appropriate) and coded information relating to the breach such as the date of breach, size, and cause of
the breach, types of information compromised, and all causes of action. We also identified whether any
dispositive motions were filed, and coded the disposition of the case. Settlement information (such as
actual confirmation of a settlement, and amounts of any damage awards) was obtained either from the
docket filings, or from directly contacting the litigating attorneys.

Data Generating Process

Data breach and lawsuit data are generated from the processes shown in Figure 1.

Stage 1: Reported and unreported breaches. As mentioned, for the purpose of this manuscript, a
“data breach” is defined as the unauthorized disclosure of personal information. From this population of
events only a subset will become public knowledge and “reported” by the Datalossdb clearinghouse.
Specifically, the only breaches that are included in this clearinghouse are those relating to social security
numbers, financial/banking information, credit card numbers, or medical information, and where the
number of records compromised exceeds 10. This group has been systematically collecting data breach
information since 2005.

Stage 2: Non-litigated, state-litigated, and federally-litigated data breaches. Stage 2 describes
three separate outcomes from the sample of reported breaches: non-litigated, federally-litigated, or state-
litigated. Because our key research questions relate to federal policy solutions to resolving the
externalities caused by data breaches, our empirical focus compares federally-litigated breaches with non-
federally-litigated breaches (i.e. both state- and non-litigated breaches). It is important to note that by

---

5 Note that the sample of “unreported breaches” (the dotted line from Stage 1 to Stage 2) also contains observations which would
be non-litigated, federally-litigated, or state-litigated. However, when addressing our first research question (“Which data breaches
are litigated?”), we do not include these observations.


7 Arbitration is one further category of outcome that may exist. In these cases, plaintiffs, as a result of enjoying a firm’s good or
service, are contractually bound to resolve any legal dispute through arbitration, rather than civil court. However, we are unaware of
any arbitrations in which privacy rights have been adjudicated.
pooling state- and non-litigated breaches we are still able to obtain unbiased estimates of federal lawsuits resulting from reported data breaches.\(^8\)

Stage 3: Federal lawsuits observed from Westlaw. For Stage 3, we obtained a sample of federal lawsuits through Westlaw using a systematic search strategy designed to identify the largest collection of data breach lawsuits practical, and then manually edited the list of suits matching our research question. Investigations by researchers have concluded that the Westlaw Pleadings database (used in this analysis), “covers or nearly covers the universe of federal claims [as it related to veil piercing lawsuits]” and that it “was designed to collect all federal complaints since 2000 that lawyers litigating commercial cases would have a plausible interest in learning about.” (Boyd and Hoffman, ND). Therefore, we do not believe that the use of Westlaw would pose any significant selection bias for our analysis.

It is relevant to also mention that the sample of unreported breaches may result in no federal or state litigation, although - for clarity - only the path to federally-litigated breaches is drawn in Figure 1 (these data are included for the purpose of our second research question: “Which data breach lawsuits settle?”).

**Which Data Breaches are Litigated in Federal Court?**

**Hypotheses**

Cooter and Rubinfeld (1989) examine prior theoretical models of litigation to create a unified framework for legal disputes. They present an analytical foundation describing the tensions faced by injurer and victim (defendant and plaintiff) at each stage of a dispute. First, when deciding whether or not to prevent an accident, an injurer balances the (marginal) cost of care with the (marginal) cost of an accident. Then, when deciding whether or not to sue, a plaintiff compares the cost of litigation with the expected benefit from an award. Finally, when deciding whether to settle or proceed to trial, both plaintiff and defendant balance their expected costs of litigation with the outcome from trial. This section is concerned with the second stage (the alleged victim’s decision to file suit), which is increasing in both the probability of success and magnitude of award (her expected gain). Below, we adapt these conditions to data breach litigation to construct appropriate hypotheses.

First, we consider the magnitude of a potential award. Given that most data breach lawsuits are class actions, the magnitude of a plaintiff’s award becomes a function of the size of the class, which is proportional by the number of records compromised in the data breach. If it is true that class action lawsuits are, in general, driven by class action plaintiffs’ attorneys, it follows that the larger the data breach, the greater the potential fee award to the attorney, and the greater the incentive to bring and litigate the suit.\(^9\) *Therefore, the probability of a lawsuit is positively correlated with the number of records lost (H1a).*

Next, the probability of a favorable outcome is multifaceted. Among other things, it is a function of whether an alleged harm can be attributed directly to the breach, the cause of the breach, and the types of information lost.

Plaintiffs in many data breach lawsuits seek relief for harms such as actual financial loss from identity theft, emotional distress, costs of credit monitoring, and anticipated future losses. However, a critical factor affecting the success of a lawsuit is the presence of a cognizable harm for which the law could provide a remedy. In the context of data breach litigation, this is manifested by whether or not the plaintiff can allege (though would not yet have to prove) financial harm. Moreover, plaintiff harm (loss) is also a function of whether the breached firm provided any initial compensation immediately following the breach and before litigation. This redress is commonly offered in the form of credit monitoring or identity theft insurance. Full compensation for any loss will decrease plaintiffs’ legal remedies. *Therefore, the*

---

\(^8\) Alternatively, had we complete data on all three outcomes, one might choose to estimate a multinomial logit model in order to separately estimate marginal effects on federal- versus state-litigated breaches. Or, one might pool state and federal suits together in order to draw inferences about all litigated breaches. However, because our topic of interest is primarily federal policy matters, we pool all non-federally litigated outcomes (that is, state and non-litigated breaches).

\(^9\) It is not the purpose of this research to address the motivations of attorneys, but merely to understand and apply relevant behavior in forming reasonable hypotheses. Conversations with class action plaintiffs attorneys confirm that while it is true that attorneys do seek plaintiffs, plaintiffs also seek attorneys for class action litigation.
The probability of a lawsuit is positively correlated with the presence of actual harm, and negatively correlated with credit monitoring (H1b).

The legal merits matter. In the context of data breaches, a plaintiff’s case is strengthened by her ability to prove that the defendant had a legal duty to protect their personal information, and somehow failed in that duty. This could occur in two different ways.

The first manner relates to the cause of the breach, which typically occurs in one of three ways: improper disclosure or disposal of personal information (e.g. tossing tax records in a dumpster); a computer hack (e.g. computer-based theft of information); loss or theft of hardware (e.g. petty theft of computer hardware that happens to contain personal information). Of these methods, we consider that the first cause (the careless handling of personal information) may provide the strongest legal argument, because it involves the negligent behavior on the part of the data custodian, as opposed to the misfortune of petty theft. Therefore, lawsuits are more likely to occur from breaches caused by improper disclosure of information, relative to the computer hack, or loss of hardware (H1c).

The second manner relates to the types of information compromised. It is reasonable to consider that the greater the legal duty to protect certain information (typically enforced through statute), the greater the probability of a favorable outcome. For instance, organizations using medical and financial data are governed by a regulatory environment requiring the enhanced protection of such data. The Health Information Portability and Accounting Act (HIPAA) requires patient consent before the disclosure of medical information between health agencies. The Gramm-Leach-Bliley Act (GLBA) and Fair Credit Reporting Act (FCRA) require greater security controls protecting an individual’s credit data. In addition, many state and federal laws require the proper disposal of social security numbers (Dickey et al., 2011) and the storage and transmission of credit card data is also protected through contractual agreements by the credit card companies under the Payment Card Industry Data Security Standard (PCI-DSS). Therefore, the probability of a lawsuit is positively correlated with the compromise of personal information requiring a heightened level of protection, such as social security numbers, financial, credit card and medical data (H1d).

Descriptive Statistics

Our final sample of DataLossdb data consists of 1,772 US data breach observations, of which only 65 (3.7%) were litigated in federal court. Figure 2 compares the number of reported data breaches with the number of federally-litigated breaches during the period 2005 to 2010. In the left panel, lawsuits are scaled according to the left axis (0-16), while reported breaches are scaled according to the right axis (0-600). The right panel shows the ratio of filed lawsuits to the number of breaches reported in that year (i.e., the portion of federally-litigated breaches over time). The right panel shows that, in 2005, the proportion of federal lawsuits was about 10%. However, since 2005, the proportion of federal lawsuits appears to be declining slightly, reaching around 3% in 2010.

Figure 2. Reported breaches vs. known lawsuits.

Note that we employ the general categories used in the DataLoss clearinghouse and that these categories are not mutually exclusive: a data breach can compromise one or more types of data.
Figure 3 compares breaches that were and were not federally-litigated as a function of the types of personal information compromised. Note that a single breach may result in the compromise of multiple types of personal information.

![Figure 3. Types of personal information compromised.](image)

Breaches involving financial data (FIN) and credit card numbers (CCN) are more likely to be litigated in federal court, which provides some support for H1c. Social security numbers (SSN), on the other hand, compromised about 78\% of non-litigated breaches, though only 58\% of litigated breaches. Medical data (MED) appear to be equally represented in federally-litigated and non-federally-litigated breaches.

**Estimating Model**

To test hypotheses H1a-H1d, we estimate a binary outcome model predicting the probability that a reported data breach will result in a federal lawsuit,\(^\text{11}\)

\[
lawsuit = c_0 + \text{ActualHarm} + \text{CreditMonitoring}_i + \text{Cause}_i + \text{ProtectedPII}_i + \text{Controls}_i + \epsilon_i
\]

(1)

where \(lawsuit\) is a binary variable that takes the value 1 if a reported breach, \(i\), results in a federal lawsuit, and 0 otherwise.\(^\text{12}\) Although we cannot determine with absolute certainty whether financial loss had occurred following a data breach, we can proxy for this by observing any evidence from news reports following the breach. Therefore, \(ActualHarm\) is coded as 1 if we observe any evidence of financial loss due to the breach, and 0 otherwise.\(^\text{13}\) \(CreditMonitoring\) is a dummy variable coded as 1 if there was any evidence that the breached firm provided any sort of credit monitoring or identity theft insurance to the individuals following the breach.\(^\text{14}\) \(Cause\) is a vector of mutually exclusive and completely exhaustive dummies reflecting the cause of the data breach: improper disclosure or disposal, computer hack or

---

\(^\text{11}\) Eq. 1 is shown as a linear probability model for clarity only. Actual regressions are estimated using logit. Also note that we limit inferences to predictions of the probability of a known federal lawsuit conditional on a reported data breach.

\(^\text{12}\) Note again that this coding inherently pools state-litigated and non-litigated breaches, thereby ensuring that estimates of federal lawsuits from reported breaches are unbiased.

\(^\text{13}\) Of the 1772 data breaches, we were unable to find news reports for 83 of them. In the absence of evidence, we took the most conservative approach and coded these breaches as not causing actual harm. We then performed a robustness check by considering that all 83 observations did cause actual harm. All estimates maintain qualitative magnitude and significance except for \(ActualHarm\) which reduces in magnitude by one third and therefore loses statistical significance. One may also be concerned that plaintiffs may wait many years following a breach before filing suit, however we do not find evidence of this. In a sample of 146 single-suit breaches, 78\% were filed within one year, and 87\% were filed within two years of public notification.

\(^\text{14}\) This information was obtained from breach disclosure notices obtained by the Datalossdb clearinghouse, or through news reports, when available. Given that perfect information is not always available, we code this variable equal to 1 only when there is actual evidence of redress. As a result, this variable is likely an under-estimate of the true frequency.
lost/stolen hardware.\textsuperscript{15} \textit{ProtectedPII} is a vector of dummies representing types of personally identifiable information (PII) should require a heightened level of protection, as described in the hypothesis: social security number, medical, financial, credit card). \textit{Controls} represents a vector of controls for all other data types (email address, name/address, etc), industry of the breached firm, whether the firm was a non-profit or publicly traded, and year dummies (2005 to 2010). $\epsilon_i$ is the random error term, assumed to be independent of the observed covariates. Descriptive statistics for the variables used in Eq. 1 are shown in Table 3.

Results

The results of Eq. 1 are presented in Table 1 and reflect the average marginal effects of the explanatory variables on the probability of lawsuit estimated using a logit regression.\textsuperscript{16} Model 1 presents just the variables of interest from H1a-H1d and includes only Year controls, whereas Model 2 includes all data types. Models 3a and 3b control for industry variables; they are based on the same estimating equation, but Model 3b presents the results as odds ratios.

The results are robust across all models, with the third model – which controls for all variables - providing the better fit for the data and generally more conservative estimates. Though not shown, results are also robust to the exclusion of individual years 2005-2010, and to probit models. Further discussions therefore focus on results from Model 3a.

In regard to the effect of the size of the breach on probability of lawsuit, our results suggest that a 10-fold increase in the number of compromised records increases the average probability of lawsuit by 8% (from 3.7% to 11.7%), a statistically significant amount (at the 1% level), which supports H1a.\textsuperscript{17}

Supporting H1b, the presence of actual (financial) loss is associated with a 2.5% increase in the probability of litigation (though, only significant at the 10% level), while the presence of credit monitoring is associated with a 3.7% decrease in probability of litigation (significant at the 1% level). Described in terms of odds-ratios (Model 3b), these results suggest that the odds of a firm being sued are 3.5 times greater when individuals suffer actual (financial) harm, but 6 times lower (1/0.152) when they provide free credit monitoring following a breach. While credit monitoring is widely touted by as a best practice following a data breach and, indeed, is included as part of a recent federal data security bill (HR2221), we provide the first statistical evidence to substantiate the practice’s value in reducing an organization’s ex post liability costs.

Next, we examine the relative odds of a lawsuit occurring given the different cause of the data breach (unauthorized disclosure, hack, or lost/stolen). Our results suggest that the odds of a firm being sued due to the unauthorized disclosure/disposal of consumer information are 3 times greater, relative to breaches caused by lost/stolen data (significant at the 5% level), supporting H1c. These results suggest that individuals are much more likely to punish firms when the firm is thought to have behaved negligently with consumer information, relative to the firm being the unfortunate victim of computer hardware theft.

Among all types of personally identifiable information (PII) requiring greater protection, we find that only the compromise of financial data is significantly correlated with the probability of lawsuit: the compromise of financial data increased the probability of lawsuit 5.1% (significant at the 1% level), which provides only partial support for H1d. That is, the odds of a firm being sued are 6 times greater when the breach involved the loss of financial information.

\textsuperscript{15} As is customary with categorical variables, we will omit one of these from the regression analysis. Given that the selection is arbitrary, we omit “lost/stolen.”

\textsuperscript{16} Note that the marginal effects for logit models are nonlinear functions of the parameter estimates, and so the effect of a regressor on the probability of lawsuit can either be presented as the effect for the “average observation” (i.e. marginal effect computed at the sample mean of the regressors) or, the “average effect” (i.e. computing the marginal effect for all observations and taking the average). We believe the second approach is more appropriate for our model because: 1) we avoid the confusion of subjectively determining the value of the regressor at which to compute the marginal effect, as in the case of the logged regressor, and 2) given that most explanatory variables are dummies, we do not need to justify having to calculate the marginal effect at a sample mean of a binary regressor.

\textsuperscript{17} A 10 fold increase represents a change of 900%, or $0.009*9 = 0.081$ or 8.1\%.
Overall, we find that our hypotheses support theoretical models of litigation. In this arena, dominated by class-action practice, parties appear to behave in a rational and wealth-maximizing manner. In the context of data breaches, this translates to a higher probability of a federal lawsuit given evidence of actual financial loss, stronger claims of negligence (unauthorized disposal of information), and heightened protection of personal financial information. However, notwithstanding the statistically significant results, none were large in magnitude. That is, no marginal effect was larger than 5%. It is yet unclear whether the magnitude of these findings is, in itself, unexpected, though it does warrant further consideration.

Next, we examine the characteristics of data breach lawsuits leading to settlement.

**Which Data Breach Lawsuits Settle?**

**Hypotheses**

The previous section leveraged the theoretical analysis of dispute litigation to develop hypotheses explaining the probability of a federal data breach lawsuit. We continue that process to develop hypotheses regarding the probability of settlement once a suit has been filed.

Cooter and Rubinfeld (1989) consider that a plaintiff (and her attorney) will decide to settle when the expected gains from settlement exceed the expected gains from trial. However, the vast majority of data breach lawsuits terminate before trial, either through dismissal or by settlement. Indeed, of over 230 suits in our dataset, we observe only two instances of a plaintiff prevailing on a favorable ruling by a judge or jury. Therefore, we can simplify the theoretical model by stating that a plaintiff (and her attorney) will settle when the expected benefits from a settlement award exceed the cost of further litigation. We now adapt this theory to data breach litigation by examining conditions that would increase either the probability or magnitude of settlement.
The recognition of the legal merits or "case strength" of a lawsuit has been the topic of much analysis in legal scholarship (see, generally, Boyd and Hoffman, ND, and Eisenberg and Lanvers, 2009; and see Johnson et al., 2007, Cox et al., 2008, and Choi, 2007, in regard to securities class action litigation). Data breach lawsuits are often dismissed because of lack of identity theft following the breach (GAO, 2007). However, there are cases when plaintiffs do suffer actual harm and are therefore able to overcome this procedural obstacle and obtain settlement. Hence, we consider that in the context of data breach lawsuits the presence of "actual harm" represents an appropriate measure of a meritorious legal claim that should affect the probability of settlement. Therefore, the probability of settlement is positively correlated with lawsuits in which the plaintiff is able to demonstrate actual harm (H2a).

A second factor which may affect the magnitude of the settlement award is whether, in class action lawsuits, the class achieves certification. Class certification represents the difference between damages potentially awarded to only a few named plaintiffs, versus thousands or millions of plaintiffs. Indeed, "class certification stands not as a mere judicial byway on the road toward full-fledged trial on the merits but, almost invariably, as the last significant judicial checkpoint on the road toward settlement" (Nagareda, 2010, p152). Therefore, the probability of settlement is positively correlated with achieving class certification (H2b).

A final driver potentially affecting the magnitude of settlement is statutory damages. Plaintiffs bring many kinds of common law claims (e.g. negligence, breach of contract) and statutory causes of action. For example, the Computer Fraud and Abuse Act (CFAA), the Fair Credit Reporting Act, and Electronic Communications Privacy Act. A defining characteristic of these Acts is their mere violation can justify plaintiff relief through statutory damages. For example, the Wiretap Act allows recovery up to $100 per day or $100, whichever is greater; the CFAA allows statutory damages of $5000 per incident (record compromised). Hence, we consider that defendants may be more likely to settle when complaints include causes of action with statutory damages. The reasons are twofold. First, these allegations shift the burden from the plaintiff having to demonstrate harm to the defendant having to prove that they did not violate the law, increasing the defendant's cost of litigation. Indeed, "the only real significant liability threat to those companies sustaining a data breach is the advent of statutory damages — damages that would ensue with or without any showing of real harm to a plaintiff" (Paray, 2011). Second, there may be a saliency effect when the defendant is forced to consider the potentially massive damage award that is the product of the statutory damages and the size of the class. Therefore, the probability of settlement is positively correlated with lawsuits in which the plaintiff seeks statutory damages (H2c).

Descriptive Statistics

To address our second research question, we relax the restrictions imposed in Section 4 and employ our full set of federal data breach lawsuits. Note that this dataset is more comprehensive than that used in Section 4, in that it includes all federally-ligated breaches (though we omit pending and public action suits). The resulting dataset of 164 observations consists of lawsuits that terminated either by settlement (n=86) or dismissal (n=78).

Figure 4 examines the proportion of cases in which plaintiffs were able to show actual damage (H2a), where the case achieved class certification (H2b), and where the plaintiff sought statutory damages (H2c). Note that in the following figures, percentages sum to 100% in each adjacent column pair.

The top two pair-wise comparisons illustrate a similar result: the majority of cases that allege actual harm or achieved class certification, settled. That is, of the cases that alleged actual harm (n=28), 71% of them settled, whereas only 49% of them without actual harm (n=135) settled. Similarly, of the cases that achieved class certification, 85% settled, whereas when the class was not certified, only 48% settled. The bottom panel, on the other hand, is more balanced. Of the cases that include causes of action with statutory damages, 59% settled, and only about 45% otherwise. Again, note that these figures reflect data from all years, and that the patterns presented in both panels are robust across individual years.
Figure 4. Pair-wise comparisons by settlement.

**Estimating Model**

We again employ a discrete outcome model to estimate the probability of settlement,

\[
\text{settlement}_i = \alpha 0 + \text{ActualHarm}_i + \text{ClassCertified}_i + \text{StatutoryDamages}_i + \text{Controls}_i + \varepsilon_i
\]

where \text{settlement}_i is a binary outcome variable coded as 1 if the lawsuit terminated in settlement and 0 otherwise. \text{ActualHarm}_i is coded as 1 if the plaintiff’s complaint alleges an actual loss due to the breach (for instance, if the plaintiff alleges fraudulent charges on a credit card, stolen money from a checking or savings account, or other such costs incurred from criminal activity). \text{ClassCertified}_i is coded as 1 if the suit achieved class certification. \text{StatutoryDamages}_i is coded as 1 if the complaint alleged violation of a federal statute allowing for statutory damages. \text{Controls}_i is a vector of explanatory variables that includes size and cause of the breach, types of information lost, industry and circuit court controls, number of causes of action and number of times the complaint was amended, and year when the case was disposed. \varepsilon_i is the random error term, assumed to be independent of observed covariates. Descriptive statistics for the variables used in Eq. 2 are shown in Table 3.

**Results**

Table 2 presents the results of Eq. 2, reporting the average marginal effects of the explanatory variables on the probability of settlement. Model 1 includes just the variables of interest and year fixed effects, while Model 2 includes subsequent controls for Breach and Industry characteristics. Model 3a and 3b include the full set of controls and estimate the same equation, with Model 3b presenting the results as odds-ratios. Further discussions therefore focus on estimations from Model 3a.
Table 2  
Regression results Eq. (2)  

<table>
<thead>
<tr>
<th>Dep var: settled</th>
<th>Basic model (1)</th>
<th>With breach and industry controls (2)</th>
<th>Full model (3a)</th>
<th>Full model (odds-ratios, 3b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Harm</td>
<td>0.275***</td>
<td>0.310***</td>
<td>0.302**</td>
<td>9.19</td>
</tr>
<tr>
<td></td>
<td>(0.096)</td>
<td>(0.106)</td>
<td>(0.119)</td>
<td></td>
</tr>
<tr>
<td>Credit Monitoring</td>
<td>-0.041</td>
<td>-0.008</td>
<td>0.102</td>
<td>2.11</td>
</tr>
<tr>
<td></td>
<td>(0.101)</td>
<td>(0.130)</td>
<td>(0.145)</td>
<td></td>
</tr>
<tr>
<td>Class Certification</td>
<td>0.407***</td>
<td>0.327**</td>
<td>0.304***</td>
<td>9.31</td>
</tr>
<tr>
<td></td>
<td>(0.140)</td>
<td>(0.143)</td>
<td>(0.117)</td>
<td></td>
</tr>
<tr>
<td>Statutory Damages</td>
<td>0.163**</td>
<td>0.192*</td>
<td>0.097</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td>(0.078)</td>
<td>(0.103)</td>
<td>(0.098)</td>
<td></td>
</tr>
<tr>
<td>Log(records)</td>
<td>0.003</td>
<td>-0.006</td>
<td>0.959</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breach_Disclosure</td>
<td>0.085</td>
<td>0.170</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.138)</td>
<td>(0.135)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breach_Hack</td>
<td>0.243**</td>
<td>0.290***</td>
<td>9.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.122)</td>
<td>(0.111)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PII_SSN</td>
<td>0.113</td>
<td>0.078</td>
<td>1.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.101)</td>
<td>(0.108)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PII_Medical</td>
<td>0.310**</td>
<td>0.312***</td>
<td>15.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.142)</td>
<td>(0.094)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PII_Financial</td>
<td>-0.123</td>
<td>-0.072</td>
<td>0.589</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.114)</td>
<td>(0.096)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PII_Credit Card</td>
<td>-0.083</td>
<td>-0.045</td>
<td>0.715</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.118)</td>
<td>(0.109)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year Controls</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Circuit Court</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region Controls</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PII Controls</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industry Controls</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forum Controls</td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>158</td>
<td>156</td>
<td>154</td>
<td></td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-93.475853</td>
<td>-78.888117</td>
<td>-64.067586</td>
<td></td>
</tr>
<tr>
<td>Pseudo R^2</td>
<td>0.1456</td>
<td>0.2701</td>
<td>0.3991</td>
<td></td>
</tr>
</tbody>
</table>

Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

These results suggest that, after controlling for all variables, plaintiff allegations of financial harm are correlated with a 30% increase in the probability of settlement (from 52% to 68%, significant at the 1% level), supporting H2a. Similarly, the certification of a class action, as Nagareda (2010) theorizes, increases the probability of settlement by 30% (significant at the 1% level), supporting H2b. In addition to each being highly statistically significant, these estimates are also large in magnitude and therefore of strong practical significance.

On the other hand, we find that causes of action asserting a violation of a federal statute with statutory damages were not positively correlated with settlement, lending no support for H2c. This finding is somewhat surprising given that this hypothesis had a strong theoretical and practical justification: these claims can help shift the burden of proof from the plaintiff in having to demonstrate actual harm to the defendant in having to prove it did not violate the law. A possible explanation for this result could be that the novelty of federal-statute based privacy litigation made it harder for the parties to arrive upon a shared understanding of the merits.
Interestingly, while the compromise of financial data and breaches caused by improper disposal/disclosure appeared to drive litigation, the compromise of medical data and breaches caused by cyber attack appear to drive settlement. Moreover, even without actual harm or class certification, lawsuits still tend to settle about half of the time. That is, cases with merit were much more likely to settle - yet, cases without merit still settle about half of the time.

A possible explanation could be that defendants choose to settle for reasons entirely unrelated to the merits of a case. For example, they may be rationally choosing to settle to avoid further litigation costs, publicity, or distraction. Specifically, defendants may be balancing between the costs of an immediate and “certain” settlement, versus a future “uncertain” amount (that includes a settlement award with some probability in addition to legal fees). Nevertheless, a full explanation, we believe, warrants more consideration.

Discussion and Conclusion

Recent events concerning breaches of consumer personal information have prompted a flurry of lawsuits by alleged victims of identity theft. These disputes have generated considerable Congressional activity concerning the collection, use, and dissemination of personally identifiable consumer health, financial and behavioral information. But is litigation an effective solution?

Consider both the probability of data breach litigation and settlement. On one hand, the overall federal litigation rate for reported data breaches is only about 4%, which may provide comfort to firms (potential defendants) that collect personal information. On the other hand, the settlement rate for all known federally litigated breaches is much higher than one might expect (50%), which would alternatively be encouraging to plaintiffs. Moreover, if actual harm (as defined within this manuscript) is indeed an appropriate measure of case merit, then the results presented here may provide some assurance that data breach lawsuits are being appropriately disposed of, on average. That is, those cases that should settle (because of the presence of actual harm), do settle. In fact, the top left panel of Figure 4 suggests that defendants settle perhaps too often (i.e. in absence of actual financial harm, and therefore case merit).

References


Government Accountability Office. 2007. Data Breaches Are Frequent, but Evidence of Resulting Identity Theft Is Limited; However, the Full Extent Is Unknown. GAO publication GAO-07-737.


Appendix

Table 1.
Summary Statistics for Eq. (1) and Eq. (2).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Eq. (1), n = 1772</th>
<th></th>
<th>Eq. (2), n = 164</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>Log(records compromised)</td>
<td>7.91</td>
<td>2.87</td>
<td>9.58</td>
<td>5.46</td>
</tr>
<tr>
<td>Actual harm</td>
<td>0.05</td>
<td>0.21</td>
<td>0.17</td>
<td>0.38</td>
</tr>
<tr>
<td>Breach_disclosure</td>
<td>0.23</td>
<td>0.42</td>
<td>0.58</td>
<td>0.50</td>
</tr>
<tr>
<td>Breach_hack</td>
<td>0.28</td>
<td>0.45</td>
<td>0.23</td>
<td>0.42</td>
</tr>
<tr>
<td>PII(SSN)</td>
<td>0.77</td>
<td>0.42</td>
<td>0.37</td>
<td>0.48</td>
</tr>
<tr>
<td>PII_medical</td>
<td>0.12</td>
<td>0.33</td>
<td>0.09</td>
<td>0.29</td>
</tr>
<tr>
<td>PII_financial</td>
<td>0.09</td>
<td>0.28</td>
<td>0.27</td>
<td>0.45</td>
</tr>
<tr>
<td>PII_creditcard</td>
<td>0.12</td>
<td>0.32</td>
<td>0.26</td>
<td>0.44</td>
</tr>
<tr>
<td>PII_email</td>
<td>0.03</td>
<td>0.16</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>PII_nameaddress</td>
<td>0.77</td>
<td>0.42</td>
<td>0.34</td>
<td>0.47</td>
</tr>
<tr>
<td>PII_datedbirth</td>
<td>0.16</td>
<td>0.37</td>
<td>0.15</td>
<td>0.36</td>
</tr>
<tr>
<td>Ind_business</td>
<td>0.27</td>
<td>0.44</td>
<td>0.49</td>
<td>0.50</td>
</tr>
<tr>
<td>Ind_education</td>
<td>0.28</td>
<td>0.45</td>
<td>0.02</td>
<td>0.16</td>
</tr>
<tr>
<td>Ind_financial</td>
<td>0.12</td>
<td>0.33</td>
<td>0.28</td>
<td>0.45</td>
</tr>
<tr>
<td>Ind_government</td>
<td>0.18</td>
<td>0.38</td>
<td>0.12</td>
<td>0.32</td>
</tr>
<tr>
<td>Non-profit</td>
<td>0.03</td>
<td>0.16</td>
<td>0.18</td>
<td>0.38</td>
</tr>
<tr>
<td>Publicly traded</td>
<td>0.12</td>
<td>0.32</td>
<td>0.41</td>
<td>0.49</td>
</tr>
<tr>
<td>Class action suits</td>
<td>0.76</td>
<td></td>
<td></td>
<td>0.43</td>
</tr>
<tr>
<td>Class certification</td>
<td>0.12</td>
<td></td>
<td></td>
<td>0.33</td>
</tr>
<tr>
<td>Statutory damages</td>
<td>0.54</td>
<td></td>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td>Multisuit cases</td>
<td>0.18</td>
<td></td>
<td></td>
<td>0.38</td>
</tr>
<tr>
<td>Removed</td>
<td>0.14</td>
<td></td>
<td></td>
<td>0.36</td>
</tr>
<tr>
<td>Female judge</td>
<td>0.24</td>
<td></td>
<td></td>
<td>0.43</td>
</tr>
<tr>
<td>Settled</td>
<td>0.52</td>
<td></td>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td>Standing</td>
<td>0.08</td>
<td></td>
<td></td>
<td>0.27</td>
</tr>
<tr>
<td>Log(employees)</td>
<td>8.73</td>
<td></td>
<td></td>
<td>2.80</td>
</tr>
</tbody>
</table>
Hand in Glove, or Square Peg, Round Hole: An Exploration of Social Network Analysis Applied to Online Community Trace Data

Kimberly Glasgow
Johns Hopkins University Applied Physics Laboratory
University of Maryland
kimberly.glasgow@jhuapl.edu

Abstract

Online communities have been studied using social network analysis. These studies often utilize online community data extracted from digital trace data or user behavior observed in the system used by the community. However, the relationship between the networks depicted in the trace data and the social networks reported or observed by community members themselves is not well understood. As a result, the implications for interpretation of social network measures in this context are unknown. This is a fundamental gap in our understanding of and ability to analyze online community. This study analyzes empirical data for both trace networks and respondent-reported networks, and compares these networks across a number of social network measures for a salient online community. Significant differences between trace networks and reported social networks are found. These differences are observed at multiple levels of analysis, from the individual to the whole network.

Keywords: social networks, digital trace data, online community

Introduction

Online communities, particularly those supported by technology platforms such as social media or social networking platforms, are a relatively new phenomena, dating back only decades. Yet people are engaging with and participating in online communities at a growing rate, and the majority of adult Americans now participate in social networking sites (Zichuhr & Madden, 2011). Many research methods and approaches have been applied to improve understanding of the nature and implications of online community, including social network analysis (Preece & Maloney-Krichmar, 2011).

Social network analysis (SNA) as a field of scientific inquiry, focused primarily on human social networks, emerged over 70 years ago. The field has a fundamental focus on the ties that connect individuals - bonds of kinship, friendship, affection, communication, social support, and of many other types. Typically, these ties are neither randomly formed nor completely predetermined. Instead they reflect individual choices made within a larger social context. Some social ties reflect internal states or emotions felt towards others. Data on these relationships can be gathered through the self-reports of network members. Other relationships may be easily observable in the public sphere, or may be documented in official records (such as marriage ties).

In short, SNA is a relationally driven approach to understanding actors (individuals), groups, and social phenomena. It emphasizes that network structure and actor position within the network are important objects of study (Borgatti, Mehra, Brass, & Labianca, 2009). Networks may serve as a conduit for opportunities for network members, or they may embed an actor in a web of connections that constrain the actor's actions. Networks have emergent properties, which can driven by underlying mechanisms, such as the social processes involved in individual choices to form or dissolve social ties with others.

Acknowledgement: The author would like to acknowledge Dr. Bo Xie and Dr. Jordan Boyd-Graber, University of Maryland, School of Information Studies, for their contributions and support.

Glasgow, K. (2013). Hand in glove, or square peg, round hole: An exploration of social network analysis applied to online community trace data. iConference 2013 Proceedings (pp. 138-154). doi:10.9776/13166

Copyright is held by the author.
Numerous social network measures, grounded in social science, have been developed to quantitatively characterize social networks, and the individuals and subgroups they are composed of. The study of social networks has contributed to our knowledge in areas as diverse as diffusion of innovation (Valente, 1996), health and disease (Christakis & Fowler, 2007), social capital (Wellman, Haase, Witte, & Hampton, 2001), and organizational structure and performance (Borgatti & Foster, 2003). An ongoing challenge for SNA has been gathering appropriate, accurate and sufficient network data (Marsden, 1990).

Technologies used to support online communities typically capture substantial amounts of data about user interactions, recording the "digital trace" of each user's behavior within the online community. These data are abundant, time-stamped, and free of human memory biases or recall error (Howison, Crowston, & Wiggins, 2011). These data may contain numerous links. All these features would appear to make digital trace data ideal for social network analysis. However, these data are system-generated, and were generated to support system requirements and processes. They may differ substantially from data that individuals would report about their personal social networks, and thus may not accurately reflect a meaningful social reality that would be recognized or accepted by network participants.

Making sense of networks in online community requires more than taking trace data at face value. It requires careful consideration of the extent to which the trace data mirrors phenomena of interest. To do so, we must improve our understanding of how the ties and networks derived from trace data relate to networks of personal ties, those ties that are recognized by and continue at the discretion of the people involved (Burt, 1997). Such understanding could underpin extending social network analysis in a valid way to this area. This work can illuminate understanding of where and how existing social network analysis methods and measures can be applied to trace data, and where new phenomena arising from online community and its trace data require new approaches. As the quantity of digital trace data and interest in it explodes, it becomes ever more imperative to better understand the phenomena.

This study will compare digital trace data and social network data gathered via survey for a specific community, “Terra.” The Terra community possesses a set of online and offline characteristics that make it particularly relevant for this endeavor.

Background

The study of social networks has a history spanning the better part of a century, providing perspectives on social phenomena and contributing to disciplines ranging from economics to psychology to organizational behavior (Borgatti et al., 2009).

The field of social network analysis evolved in a period when gathering network data could be an intensive, even arduous process, typically involving surveys and interviews of subjects. Additionally, the data may contain inaccuracies. Informant accuracy in self-report, in circumstances where actual behavioral observations were also made, has been observed to be imperfect. Memory fades, and informants may overreport or underreport behaviors and ties with alters (relationship partners). Yet for data on beliefs and attitudes, there may be no reasonable alternative (Bernard, Killworth, Kronenfeld, & Sailer, 1984).

The rise of the Internet, online community, and social media has dramatically changed both the landscape for social interaction, and the amount of data potentially available relating to social phenomena. As massive amounts of online data capturing, or at least reflecting, ever-larger portions of human activity become available, new opportunities will emerge (Watts, 2007). In this context, boyd characterizes the technology affordances of new means of computer-mediated communication, particularly social networking sites, and important areas of research including networks and network structure and online/offline connection (boyd & Ellison, 2008).

The potential impact of rising use of the Internet, and growth of online communities and social networking sites on society has been an issue of broad societal concern. There is evidence that social circles of Americans have been contracting over the past decades (McPherson, Smith-Lovin, & Brashears, 2006), and that community participation has been decreasing (Putnam, 2001). Such changes have been linked to time spent on line supplanting personal interaction. Other research finds positive change in social connectivity in the same time frames. Wellman examines changes to social capital for Internet users, and finds an array of positive effects, including increased participation in communities (Wellman et al., 2001). And consistent with Granovetter's strength of weak ties theory, participation in
online community would appear to open doors for participants, through exposure to others online who are not frequent, strong, and close partners (Granovetter, 1973).

Yet, limitations of theory and of methods will be encountered, to say nothing of issues of data access and individual privacy (Lazer et al., 2009). Initial efforts to directly apply social network measures in these contexts may not prove particularly useful. Freeman considers closely the problems posed when social network measures are employed by those lacking a social science background (such as physicists), or applied to network data that is not social in nature. He is optimistic that these hurdles can be overcome, particularly through cross-disciplinary communication and collaboration (Freeman, 2008).

Issues surrounding the validity of digital trace data for social network analysis have recently been raised (Howison et al., 2011), including issues of reliability, and mismatches of constructs and measures to trace data itself. The issues are explored through a review of relevant literature. The current study provides empirical data from both digital traces and survey for a community whose members simultaneously participated in both online and face-to-face interactions towards common goals. This near-perfect alignment of online and offline community, with concurrent, focused interaction, captured both through digital traces and participant report provides a unique opportunity to more fully understand the relationship between these networks.

Research Questions

Research questions to be examined include:

- How do social ties observed or inferred from online communities' digital trace data compare to social ties based on interactions in the physical world? How are they distinct?
- How do the networks that emerge from these ties relate? Are there patterns of similarity or difference?
- How can we evaluate the nature and relevance of ties created in trace networks? How do they matter? Can we align them with other types of observed or reported social ties?

Methods

This section will describe the community being studied, the data gathered and instruments used, the data analysis methods, and the tools used.

The Community

Rather than an online-community organized around a hobby or interest, the members of the community examined in this study ["Terra"] are professionals engaged in scientific research, policy, or clinical practice involving global climate and climate change issues, or public health. The goals of the community are to promote collaboration and knowledge sharing across disciplines relating to climate change. A website for the community states it is...focused on climate disruption and its effects on society. Its goals are to forge connections between, and support collaboration and communication among scientists and policy makers interested in climate change, public health, and/or national security. [It] consists of an online presence, hosted on a website, and a series of workshops, supported by that website. [source anonymized for the privacy of the community]

This community was explicitly designed to have an online and offline component, with both aspects of the community oriented around precisely the same goals and issues regarding climate change and public health. Members were expected to engage face to face during workshops, and also to participate online. In fact, during the course of a workshop, the members of the Terra community are simultaneously engaging in both online and offline interactions. Activities in workshop sessions included brainstorming, identifying and prioritizing issues, and problem solving.

All members of the Terra community had accounts and profiles on the online Terra community platform. All attended a 3-day workshop on climate change and public health. The 3-day workshop consisted of 13 half-day topical sessions, focused on issues such as food and water security, sea level
rise and extreme weather, infectious disease and water quality, and climate modeling, a number of workshop feedback sessions, and keynote sessions. For each session, background materials had been uploaded, and online interactions were developed. Thus, a participant in a session would log into the Terra system, navigate to the session, and participate in session-related online activities, such as ranking major threats to water quality. At the same time, during the course of the session, each participant was engaging in face-to-face interaction with the other session participants and the session moderator, discussing issues or sharing insights and perspectives. Over the course of the workshop, community members also had opportunities to socialize over meals and cocktails. Ninety-two people participated in the Terra workshop. Topical sessions had 7 to 18 participants (mean 14.3, standard deviation 3.85). Workshop feedback sessions had as many as 35 participants.

Data

After the workshop ended, digital trace data was collected from the Terra community platform. Participation in each session was determined through system-generated data (such as posting or voting behavior). Additionally, any free-text comments posted for a session were data mined for mentions of session participants. This produced a 92x18 matrix, in which individuals were linked to the sessions they participated in. This network is shown in Figure 1. Green squares indicate workshop sessions. Individual community members (circles) are linked to each session they took part in. The color of the circle indicates the institution (such as the specific university, government agency, or an NGO) that the individual is affiliated with. While node labels have been hidden to reduce visual clutter in this network, the cluster of green squares at the center of the network are the workshop feedback sessions.

Figure 1. Network of session participation

Sessions during the 3-day workshop fell into two categories – the topical sessions (on issues related to climate or public health), and the feedback sessions, focused on growing and improving the community. The network in Figure 2 shows participation in topical sessions only.
These networks are both 2-mode networks, in which ties link different types of entities (people and sessions, in this case). But a 2-mode network does not in itself show relationships between individuals. For further analysis, this matrix was transformed into a 92x92 affiliation matrix in which ties between individuals represents co-participation in sessions (S.P. Borgatti & Halgin, 2011). This is the "online co-participation" network (OCP). It is an undirected network. That is, if A attended a session with B, then B also attended that session with A. Visually, A and B will be connected by a line to represent this tie.

This network of affiliations between actors based upon co-participation is shown in Figure 3. While it contains fewer nodes than the previous networks, it is much denser as each community member is connected to every other member they participated in a session with.

Figure 3. Network of online co-participation
Following the workshop, all 92 community members received a social network survey. The survey elicited information on three social ties related to knowing, learning from, or working with other members of the community. These social ties relate directly to the community goals of forging connections between, and supporting collaboration among community members. The survey questions are provided in Table 1. Respondents were presented with a list of the participants in the workshop, and indicated which alters they had each tie with.

Table 1
Social Network Survey Questions

<table>
<thead>
<tr>
<th>Relationship Name</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knew Prior (K)</td>
<td>Whom of the attendees did you know prior to the &quot;Terra&quot; workshop? (To &quot;know&quot; in this context means to know by name or by face, and to have interacted with previously in a professional or personal context.)</td>
</tr>
<tr>
<td>Learned/Engaged (L)</td>
<td>Over the course of the workshop, you may have had the opportunity to both strengthen connections with existing colleagues and to meet and engage with new people. Please indicate each new person you learned from or engaged with during the conference.</td>
</tr>
<tr>
<td>Work with in Future (W)</td>
<td>While it is too soon to be certain, of the new people you learned from or engaged with, are there any that you are fairly confident you will work with in the future?</td>
</tr>
</tbody>
</table>

Thirty-five participants completed the survey, for a response rate of nearly 40%. While the response rate is higher than that of many surveys, for purposes of comparing social networks, missing data from non-respondents would affect the findings. For this reason, most analyses were performed on a subset of the data, representing those 35 Terra members for whom we had complete data - both survey data and trace data. That is, for each of these 35 participants, we have networks that represent their co-participation in sessions, and networks that represent their responses for the Knew Prior, Learned/Engaged, or Work with in the Future ties, as well as the responses that each of the other 34 provided about them. An additional network, representing the existence of any social tie between actors (K, L, or W) was generated. This is the Any Tie (AT) network. As a union of all the social ties reported by respondents, AT provides the broadest representation of the social network reported by community members.

Organizational affiliation data is available for community members from their community profiles and workshop registration. Organizational names were standardized to a single form. Additionally, organizational affiliation was mapped to the level of institution, to adjust for individuals who listed departments or subunits within a parent institution as their organization. This is the data used for actor node color in all network visualizations.

Data Analysis Methods

Analysis methods and measures have been drawn from social network analysis, as that is the relevant framework for this study. Network visualizations are supplied for each network studied.

Network Measures

Network density reflects the number of ties that exist in the network, out of the number possible (Hanneman & Riddle, 2005)

Core-periphery structure, which capture the extent to which the network is structured with a dense "core" of well-connected actors, surrounded by a "periphery" of actors who have ties to members of the core, and few ties between each other (Borgatti & Everett, 2000)

Triad Census describes the distribution of triads (sets of 3 actors), who have particular pattern of interconnections, ranging from no ties at all (an empty triad), to having all 6 possible ties (a complete triad). The triad census captures structure and connectivity at a local level, and provides some insight as to whether phenomena such as reciprocity and transitivity happen in the network (Wasserman & Faust, 1994).
Degree centrality assesses how prominent, or "central" an actor is, based on how many ties the actor has to others in the network (Freeman, 1979). Indegree captures how many ties come in to an actor, while outdegree measures how many ties originate with the actor.

Tools

Ucinet (Borgatti, Everett, & Freeman, 2002) and NetDraw (Borgatti, 2002) are used for computation of network measures and statistics, and network visualization.

Results

For reasons provided in the Data section, the majority of findings discussed will be based on data from the complete 35-person network. However, survey results for the 92-person network (not all of whom responded to the survey) will be reviewed as appropriate.

Knew Prior (K)

This network reflects relationships that existed among community members prior to attending the workshop. Unsurprisingly, there is a higher likelihood for individuals to know others from their home institution, versus individuals from other organizations. A substantial percentage of community members (the isolates on the left side of the network) may not have personally known anyone else in the community prior to the workshop. However, it is likely that some apparent isolates in this network did have existing K relationships, but that these are missing from the data, because no participant in that relationship responded to the survey.

Figure 3. Knew Prior (K)
Learned/Engaged (L)

This network reflects relationships of learning from or engaging with members of the community. These new ties formed among community members provide an indication that meaningful interaction across institutional boundaries has occurred.

Figure 4. Learned From or Engaged With (L)

Work with in the Future (W)

This network is the least connected of the three. While it is speculative, based on anticipation of activities that have yet to happen, it could be validated longitudinally. The larger number of cross-institutional ties is an indicator Terra may achieve its goal to foster new collaborations.

Figure 5. Work with in Future (W)
Complete Networks

The three networks visualized below represent a subset of the previous network data, including only actors for whom complete data was available. (All 35 actors have trace data, and all 35 completed the survey.)

Figure 6. Knew Prior (K)

Figure 7. Learned From or Engaged With (L)
Figure 8. Work with in Future (W)

Figure 9. All relationships from survey
Online Co-Participation (OCP)

This network is derived from digital trace data from the Terra online community.

Figure 10. Online co-participation relationships

Network Level Measures

**Density.** Density reflects the overall cohesion of the network, in terms of sheer connectedness. Density for the survey networks is considerably lower than for online co-participation. Even the Any Tie (AT) network has only half the density of online co-participation. The difference cannot be attributed to the undirected online co-participation network having more ties due to being an affiliation network. And there is no plausible reason to expect that participants forgot or chose not to report more than half their ties to alters. The difference in density appears to indicate the OCP network does not closely resemble any reported social network for this network level property. Network densities are reported in Table 2.

Table 2

<table>
<thead>
<tr>
<th>Network</th>
<th>Density</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knew Prior</td>
<td>0.131</td>
<td>156</td>
</tr>
<tr>
<td>Learned/Engaged</td>
<td>0.137</td>
<td>163</td>
</tr>
<tr>
<td>Work with in Future</td>
<td>0.082</td>
<td>97</td>
</tr>
<tr>
<td>Any tie</td>
<td>0.250</td>
<td>298</td>
</tr>
<tr>
<td>Online co-participation</td>
<td>0.563</td>
<td>670</td>
</tr>
</tbody>
</table>
Core-periphery structure. The degree to which a network contains a distinct core surrounded by a periphery of less-connected nodes is another aspect of overall network structure that can be used to compare networks. Comparison of the network consisting of any reported social tie with the online co-participation network reveals a considerable distinction. While both networks can be fit to a categorical core-periphery structure to a reasonable degree, they differ considerably in the size of the core, the size of the periphery, and the within-core density. The online co-participation network has a core that comprises two thirds of total network, with near perfect internal connectivity, while the AT network is half that size, and notably less well connected. Thus, the online co-participation network portrays a much more closely connected social world with a much more expansive core.

Table 3
Core-periphery structure

<table>
<thead>
<tr>
<th>Relation</th>
<th>Core Size</th>
<th>Periphery Size</th>
<th>Within-Core Density</th>
<th>Final CP Model Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Tie Online co-participation</td>
<td>12</td>
<td>23</td>
<td>0.742</td>
<td>0.628</td>
</tr>
</tbody>
</table>

Subgroup Measures

Triad Census. In additional to comparing networks at the highest level of abstraction, insight can be gained by examining networks in terms of the substructures they are composed of. Triads within a network can be examined for evidence of phenomena such as tendencies toward reciprocity or transitivity (A→B, B→C, C→A). Since online co-participation is an undirected graph and cannot contain any asymmetric triads, a comparison of the possible triads for both online co-participation and the survey networks is presented. For all cases, online co-participation has far more triads than any network based on reported social ties. For the complete triad, in which all 3 actors are perfectly connected to each other (A→B, B→C, C→A), OCP has nearly 20 times as many triads.

Table 4
Triad Census for all Networks

<table>
<thead>
<tr>
<th>Triad</th>
<th>Knew Prior</th>
<th>Learned Engaged</th>
<th>Work with in Future</th>
<th>Any Tie</th>
<th>Online co-participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-leg</td>
<td>1507</td>
<td>341</td>
<td>291</td>
<td>1125</td>
<td>2359</td>
</tr>
<tr>
<td>Complete</td>
<td>79</td>
<td>2</td>
<td>3</td>
<td>111</td>
<td>2036</td>
</tr>
<tr>
<td>Two-leg</td>
<td>111</td>
<td>27</td>
<td>25</td>
<td>235</td>
<td>1294</td>
</tr>
</tbody>
</table>

Node Measures

Degree Centrality. Degree centrality for the full 92-person shows 5 actors with scores of zero for online co-participation (inDegreeO in Figure 11), which indicates non-existent influence within a network. However, four of the 5 played key roles in the Terra workshop. These include keynote speakers, session moderators, and workshop organizers. And these 4 fall into the top third of influential actors, based on indegreeL (number of conference participants who said they learned from or engaged with them at the conference). This may reflect a limitation of the trace data for this community. Actors who played key roles or performed certain functions during the workshop, such as keynote speakers and session moderators, due to their leadership positions, appeared to participate less. Since they were leading the sessions, they typically were not voting or commenting within them. Their participation in the session could be "invisible" from the perspective of the digital trace for the session. At most, they had browsed to the session page, but that was insufficient user behavior to leave a "participation" trace in the data.
Considering just the 35-person network for which complete data is available, we can compare centrality from social ties reported in the survey (A) with centrality from the digital trace data. Again, the two values diverge. Outdegree for the digital trace data (OutDegO), the number of alters for each actor hovers around 25 for most of the actors, and significantly exceeds the number of alters reported by respondents. The anomaly noted previously, in which an actor prominent in the survey network has a centrality score of zero for online co-participation is again observed. The fifth most central actor, who reports ties to half the network members, and is both a workshop speaker and a driving force in organizing and creating the Terra community, has no social ties to anyone, based on digital trace data.

**Figure 11. Comparison of indegree centrality for Learning/Engagement and OCP networks**

![Comparison of indegree centrality for Learning/Engagement and OCP networks](image)

**Figure 12. Number of Alters**

![Number of Alters](image)
Correlation of Degree Centrality

The values for OutdegreeK and IndegreeK correlate to 95% ($R^2 = .90$). This indicates respondents are able to accurately and consistently report their perceived relationships, and are in agreement about knowing each other. Lower correlations between OutdegreeL and IndegreeL could reflect the naturally asymmetric aspect of learning, where A can learn from B, without B also learning from A. Correlations of degree between the online co-participation (OCP) network and the survey-derived networks are consistently among the lowest seen. Even for the broadest of survey-derived networks, the AT network, correlations hover around 25% for both InDegree and OutDegree.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Online Co-participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work in Future</td>
<td>0.086</td>
</tr>
<tr>
<td>Any Tie</td>
<td>0.190</td>
</tr>
<tr>
<td>Learned/Engaged</td>
<td>0.170</td>
</tr>
<tr>
<td>Knew Prior</td>
<td>0.076</td>
</tr>
</tbody>
</table>

QAP Correlation

QAP, or quadratic assignment procedure, is a standard method for determining the association between networks (Hanneman & Riddle, 2005). The correlation of OCP with any of the survey networks is too low to support it being a reasonable approximation for any of the reported social ties.

Discussion

This study examined a community, "Terra," composed of professionals engaged in scientific research, policy, or clinical practice involving global climate and climate change issues, or public health. The goals of the community are to promote collaboration and knowledge sharing across disciplines relating to climate change. Networks representing social connection between community members were compared. Community members were surveyed to determine three social networks (Knew Prior, Learned from or Engaged with, Work with in Future). A digital trace network, Online Co-participation, was created from data stored by the online community platform.

As observed by Howison, even seemingly "perfect" trace data may prove less so in reality, and require an investment in data cleaning. This study was no exception, as it was necessary to normalize participant names and to reconcile workshop attendance. Efforts expended to clean the data warn us
that trace data may need to be examined in the context of some external source, to identify and correct any potential problems.

The digital trace data showed online participation in sessions. Participation meant using the system in a way the system could record, such as posting on a topic, commenting, ranking or voting on items posted in a session. This raises the issues of what other social relationships might co-participation most closely relates to. On the face of it, the Learning/Engagement network should be quite similar, as participants were in the same room talking about session topics while they were participating in the session online. It is possible that participants considered themselves to be engaging on an issue, and not engaging with each other. The individual identities of some the others in the room may not have been “sticky” in real life, and thus may not have been remembered by others in the survey. Perhaps the online aspect of the interactions let some people “lurk” in terms of real world interaction, so they never made a true impression on others in the group, so no social ties were formed between them, despite online activity.

The trace network was found to be clearly distinct from any the others. It contained far more ties in general, and differed in terms of network level properties and structure. It was far more cohesive, with higher density and a larger core. Despite a much higher average centrality score for network members, it paradoxically produced a set of unconnected actors who were among the most central in the reported networks. This appeared to be a consequence of actors who performed certain roles in the community, including leadership and facilitation roles, being less active in the online system.

While online co-participation (OCP) certainly paints a more egalitarian picture of the network of the Terra community (higher degree, perfect reciprocity, considerable transitivity), it may be disguising actual underlying social processes that led to the lower reported values for each of these phenomena. In fact, it is not obvious how one would align this trace network with any of the survey-derived social networks, or how to interpret the significance of online co-participation to participants.

Future Directions

This study examines one community, using one technology platform, at one point in time. Limitations arising from this, and suggestions to address them in future work follow.

The data is not longitudinal. The "work with in the future" relationship is speculative, and reflects intentions. Validating with reports of joint grant applications, collaboration on research projects, or co-authorship of papers, would be helpful. Gathering more network data when the next workshop series is held, and possibly expanding the set of social ties surveyed could provide further insight, as would gathering digital trace data longitudinally for the community as it continued to develop.

Survey response was not perfect. There may be systematic differences between those who responded and those who did not. For example, those who learned/engaged the most may also have had a higher likelihood to respond to the survey. Informant accuracy in general may have affected the results. Cognitive biases in reporting ties may exist. The survey was not conducted during the workshop, but weeks later (after IRB approval was received). Participants could have forgotten ties, or the names of alters. Not all community members completed the survey. Conducting the survey earlier, perhaps during the final workshop session, could help. Community members were identified by name and institutional affiliation in the survey. Augmenting the names with pictures of participants, as a memory jog, might improve accuracy.

Different platforms for supporting online community with different features could lead to different trace data, and thus, evidence of different kinds of ties. Exploring other platforms, or platforms with different sets of features implemented or employed could provide new insights.

Three relationships were selected for this study – Knew Prior, Learned/Engaged, and Work with in Future. They align very closely with the goals of the community to foster new knowledge sharing and collaborative relationships among individuals interested in issues of climate change, health, and public policy and are clearly relevant to understanding this online community. While selecting a small number of salient relationships may help survey response rate, it is also a limitation. Selecting different relationships to survey might have produced different results.

Qualitative and quantitative approaches can provide unique and complementary perspectives of social networks and lead to deeper understanding (Fuhse & Mützel, 2011). Augmenting this work with qualitative data would be desirable. For example, incorporation of knowledge gleaned from interviews of network members could enrich our interpretation of what each of the reported social ties means to
participants, and how they experienced online co-participation as a social and community-forming interaction potentially leading to new social ties. They might propose different social ties than those elicited in the survey for capturing connections among community members.

Transforming the original trace data directly into an affiliations network of online co-participation is a logical way to proceed. Yet it is possible that applying some other type of transformation could have produced a network with properties more similar to those of the reported social networks. Exploring other types of transformations might prove fruitful.

This is a very young community. It will certainly change. Perhaps over time and with more opportunity for face-to-face interaction, trace data might start to converge with survey data. It might serve as a leading indicator for where new participant-recognized social ties might be formed.

Only one community was studied, Terra. This community consisted of self-selected individuals motivated to collaborate and share knowledge on the complex scientific, social and policy issues relating to global climate change. They further divided into subgroups focused on a subset of topics for the majority of their interactions. Thus, they may not be representative of the majority of online communities, which are likely to have far more diffuse and heterogeneous compositions. While the focus of this community provides unique opportunities, it limits generalizability. Studying more communities with different goals, sizes, membership compositions, ages, and so on would broaden our understanding.

Conclusions

The Terra community provides a unique opportunity to examine issues of personal social networks and digital trace data for an online community. Substantial differences between networks based in these types of data for the same community are observed.

The findings have implications for future work on trace data validity. For this community, the digital trace network cannot serve as a proxy for a network of personal ties, recognized by the people involved. For those social network measures examined in this study (density, core-periphery, triad census, degree centrality, QAP), digital and survey networks were dissimilar. And as these measures map onto network, subgroup, and actor-level phenomena, there was no level of network analysis for which a good correspondence was observed.

As participation in online community continues to grow, and digital trace data becomes even more abundant, the need for appropriate ways to interpret this data will not diminish. Though understanding how individuals perceive their networks and their interactions with others within an online community is important, we currently lack clear mechanisms to do so through the digital trace data, especially as it was typically designed to support system processes. Progress in this area may suggest new methods for using such digital trace data, and it may lead to improvements in technology to capture digital traces that better reflect social ties recognized by participants.

While a new field of computational social science may be unfolding before us, this study illustrates that there are still many unanswered questions about even basic correspondences between digital trace networks and personal social networks. Much work needs to be done to truly grasp the meaning and value digital trace networks can provide to our knowledge of online community. Though the fit is not yet hand in glove, the promise is powerful.

References


From Push Brooms to Prayer Books: 
Social Media and Social Networks During the London Riots

Kimberly Glasgow
Johns Hopkins University Applied Physics Laboratory, 
University of Maryland
kimberly.glasgow@jhuapl.edu

Clayton Fink
Johns Hopkins University Applied Physics Laboratory
clayton.fink@jhuapl.edu

Abstract

Social media, such as microblogging, is a powerful medium for sharing information and organizing response in times of crisis or extreme events. We propose methods to integrate topical and social information and behavior derived from social media to improve situational understanding during an extreme event. Using Twitter data from the 2011 London riots, we analyze emergent social networks directly relating to response to crisis. We construct social networks from these tweets based on talking to (directed communication), quoting (retweeting), and talking about (mentioning) behaviors. We examine networks of riot response oriented around cleanup or prayer activities. These networks differ in size, structure, and membership. We identify prominent network actors and assess their similarity. These methods may enable more effective response during disasters or other emergency events.

Keywords: social networks, social media, Twitter, crisis response, crisis informatics

Introduction

The importance of social networks in helping individuals cope with, navigate, and mitigate challenges in their environments, including crises, has been established. While the rise of social media such as microblogs has provided unique capabilities to rapidly and widely share time-critical information both to members of one’s social networks and to broader audiences, this phenomenon is poorly understood in the context of events such as civil unrest or natural disasters.

This research explores ways to identify and understand social networks in times of crisis, response, and recovery, which could contribute to improved situational understanding, crisis communications, or provision of support. Social network analysis (SNA) provides a means to understand individuals, groups, and social phenomena, from a framework in which patterns of interaction are represented as networks. The structure of the network, the positions of individuals within it, its dynamics, and any underlying social processes or mechanisms can inform our comprehension of an event and our ability to respond appropriately.

We examine the use of the microblogging service Twitter during one of the most tumultuous periods in the recent history of London. London was rocked by a series of riots in August 2011. A few days after the police shooting of an unarmed man in the ethnically diverse working class London neighborhood of Tottenham, peaceful protests were followed by an outburst of looting and arson. Five people were killed, hundreds injured, thousands arrested, and over one hundred million pounds of damage was done. Rioting spread to other urban areas in England. During these events, Twitter was used extensively in London. We analyze a corpus of tweets collected from this time. London is known as a city with a relatively high rate of Twitter usage, and many government entities have adopted Twitter as well (Panagiotopoulos and Sams 2011) to communicate with residents.

To better understand the city during a time of crisis, social media activity relating to the event must be distinguished from other activity within the larger stream. We leverage the Twitter convention of hashtags to identify relevant communications from this larger pool. Hashtags are words, phrases, acronyms, or abbreviations that can be explicitly incorporated in a tweet to signal association with the topic or meaning of the term. Hashtags are preceded by a “#”, such as #prayforlondon or #riotcleanup. Use of a hashtag makes a tweet easily discoverable by anyone interested in that tag. It is thus possible to easily follow the conversation around a topic, in contrast to following specific Twitter users.


Copyright is held by the authors.
We develop a method to categorize hashtags appearing in this data. We then select two categories of hashtags to explore whether we can find evidence of meaningful social networks in social media that are activated during extreme events. We characterize these networks in terms of network properties and prominent actors. Finally, we consider how this knowledge might be used to improve awareness, understanding, or response to these circumstances, or to gauge community reaction.

The two categories of hashtags examined in this research both involve how individuals responded to the riots. Both categories are types of responses to the damage, destruction, and pain resulting from the riots. The first category is riot cleanup. Riot cleanup is an active, public, externally visible response to the damage. Debris and rubbish are swept from the streets. Broken objects are repaired or replaced, and order is restored. The second category is prayer in response to the riots. Prayer is a more personal, internal behavior. It may be performed in private, unobservable to others. It may seek to provide a sense of comfort or control through involvement of a higher power in otherwise overwhelming circumstance. Given these differences, networks of riot response oriented around cleanup or prayer activities might be expected to differ in size, structure, and membership.

Social media can provide a window into social networks in times of crisis or disaster. The rapid reaction time of microblogging creates potential to track reactions to such events in near real time, to assess damage and suffering, and to identify individuals and networks of interest, if the relevant communications can be recognized from within the larger stream. This provides opportunities for performance improvements in emergency management, law-enforcement, and for more effective government communication and improved transparency. If such knowledge were available to everyday citizens, they might benefit from a richer understanding of information and resource flow in extreme circumstances. They might be better able to contribute to response and recovery by sharing personal knowledge or observation more effectively. Their ability to gather critical information or resources in a timely fashion could be magnified. To the degree that the social networks that emerge in times of crisis have some common core prior to the emergency, it might even be possible to proactively nurture, or plug into, these networks.

Related Work

Response to social crisis and disaster has been studied for decades from a broad range of perspectives, from social psychology and organization theory to public policy and emergency management (Quarantelli & Dynes, 1977). Beyond response by organizations with formal responsibilities during emergencies, informal, spontaneous response seems a hallmark of human behavior. For example, the phenomena of convergence by large numbers of people on the physical location of a disaster shortly after the event, often motivated by a desire to help, has long been observed (Fritz & Mathewson, 1957). With the rise of social media, related online phenomena are appearing (Hughes et al., 2008). Self-organization of individuals into informal, emergent groups to fill gaps and respond flexibly to crisis is not new (Stallings & Quarantelli, 1985), but may take new forms in a social media enabled world (Liu et al., 2008) (Starbird & Palen, 2011).

During a crisis or emergency event, the local population fills invaluable roles, from detecting early indications of an event, to providing updates as the event unfolds, to supporting recovery efforts (Starbird, Palen, Hughes, & Vieweg, 2010) (Vieweg et al. 2010). The empowering capacity of Information and Communication Technologies (ICTs) such as microblogs may expand this phenomena (Palen, et al. 2009) (Jaeger et al. 2007). Changes in online language behavior, such as blogging behavior, during crisis have already been observed (Cohn, Mehl, & Pennebaker, 2004). The new field of crisis informatics provides insights and observations on ICT use during disasters (Shkolovski, Palen, & Sutton, 2008).

Social networks are important for providing social support (Gottlieb, 1985), acquiring resources and critical information (Granovetter, 1973), and for buffering the effects of psychological trauma (Flannery, 1990) or stress, to include events such as dislocation following a natural disaster (Bland et al., 1997). Less is known about social networks in online community, though networks of empathic communication in an online discussion board have been studied (Pfeil & Zaphiris, 2009). However, research interest in this area is high despite challenges (Savage & Burrows, 2007) (Watts, 2007).

More broadly, Twitter and hashtag use in Twitter has been studied in terms of trending topics (Kwak, Lee, Park, & Moon, 2010), political polarization (Conover et al., 2011), contagion and diffusion (Romero, Meeder, & Kleinberg, 2011), and other topics.
Methods

Data Collection

Twitter data used in the research was collected from Twitter after the riots started on August 6, 2011 using Twitter’s Application Program Interface (API) for search. This API permits location-based queries and will return a sample of recent tweets from a particular location specified by a coordinate pair and a radius. We use the center coordinates of the Greater London administrative area of 51.502 latitude and -0.127 longitude, with a radius of 20 miles. To improve on the sample returned from the API, we also used the Search API to query against the users whose tweets were returned for the London query. We continued this process through September 30, 2011, covering the riots and several weeks of recovery. The data includes over 14 million tweets from hundreds of thousands of users. The Twitter API returns location metadata with tweets: either a geotag containing a latitude and longitude from users that have enabled the geotagging feature on their mobile Twitter client, or the location string from the user’s profile. The user profile location string may be a place name entered by the user or a coordinate pair populated by a mobile client. We match the location metadata against a gazetteer, looking for place name that matches or, if coordinate data, the closest populated location in the gazetteer. If we are able to match the location we use the administrative level information from the gazetteer to assign the tweet to a particular administrative district within the Greater London area. For example, a user who gives their location in their profile as Tottenham would be assigned to administrative district F3 (Haringey) and a tweet with the coordinate location or geo tag of 51.63 latitude and -0.14 longitude would be assigned to district E3 (Enfield). This step allows us to aggregate tweets by their relative location within London when there are coordinates or a place name is given that is below the level of the city. It also allows us to eliminate data that may not be from London because the place name is not resolvable because of the use of alternate spellings or nicknames (LANDANI, crackney) or because a non-place-name was given (nonyah, WithBiebs). This process does not address errors attributable to the Twitter API. For example, a user giving their location as London may be in London, England or in another London, such as London, Ontario. We have no measure of the level of this error in the data. However, since we only focus on tweets that contain content directly related to the riots and the aftermath, we argue that this error source will not significantly impact our results. Our analysis is limited to tweets that we were able to geolocate to the level of the Greater London administrative area or below.

Hashtags in Twitter

Nearly 400,000 unique hashtags appear in the overall corpus. During the week of rioting, the most common hashtag was #londonriots. This tag was used over 25,000 times that week. Many other hashtags relating to the London riots emerged spontaneously in the early days of the event. The #prayforlondon hashtag was a Twitter trending topic. Hashtags pertaining to everyday activities, such as entertainment or shopping, coexist in the data with hashtags referencing the riots. Riot-related hashtags described locations of rioting and destruction, expressed concern, or recommended responses to rioting and rioters. Hashtags relating to two aspects of response to the riots, riot cleanup activity and prayer in response to the rioting are examined in more detail.

Each of the most frequently used hashtags in the first week of the disaster was reviewed to identify a seed set of hashtags for riot cleanup. The review began with the hashtags that comprised the top half of the hashtag distribution (that is, the hashtags that accounted for 50% of total hashtag usage instances in the data). This set contained about 150 hashtags. The meaning of each these hashtags was determined through inspection of the term itself, use of online hashtag dictionaries, and through scanning the content of tweets containing the hashtag. This identified a seed set of hashtags clearly related to riot cleanup. This was used to search the remainder of the tag corpus, and a final set of 65 relevant hashtags was identified. These include #riotcleanup, #riotwombles, #londoncleanup, and #cleantottenham, and a variety of spelling and morphologic variations, such as #cleanuplondon. Tweets containing riot cleanup hashtags allow us to identify individuals in London who played a role in riot cleanup communications through Twitter.

1 http://www.geonames.org
A second seed set of prayer hashtags was constructed in the same fashion. This set contained nearly the same number of hashtags (63), though fewer tweets. The role of prayer as a potential positive mechanism for coping with and responding to crisis has been repeatedly documented (Meisenhelder, 2002). It was among the most common responses to the terrorist attacks on September 11th in the United States in a national survey of stress responses to the attacks (Schuster et al., 2001). Ninety percent of respondents reported turning to prayer, religion, or spirituality to help cope with the crisis.

**Talking to, Quoting, and Talking about within Twitter**

Twitter functionality and conventions allow a user to express more nuanced communicative behaviors than the broadcast of a short message. Users can specify an intended recipient toward whom the message is directed.

@joemcelderry91 my hometown is burning right now! #prayforlondon Please get people to RT this

Users can echo another user’s tweet by retweeting it, and credit that author.

RT @artistmakers: Getting the clean up together - Meet outside Tackle Shop, Roman Road, hackney 9am in the morning to help local shops clean up. #riotcleanup

Finally, a user can easily comment on another user by referencing that username.

**MP of the day: @HackneyAbbott. With her constituents, on the streets with those affected. Only MP in right place at right time #londonriots**

These behaviors can be considered analogous to talking to another, quoting another, or talking about another user. A sender of a tweet can easily choose any of these three behaviors when tweeting. Connecting to another in each of these ways is sociologically and semantically distinct. The patterns and networks that emerge may provide insight into roles played by individuals in these networks of crisis communication, how they are perceived by other network members, and even the robustness of the network as a whole (Borgatti, Mehra, Brass, & Labianca, 2009).

**Findings**

All tweets containing riot cleanup hashtags were used to construct a set of social networks reflecting riot cleanup communications for the three types of behaviors described above, talking to, quoting, and talking about. These tweets generated network data encompassing roughly 4000 actors and 5000 ties. The same process was used to create prayer networks. Prayer communication was less common, generating less than a third as much network data, and fewer than 1400 actors. All of the networks contained multiple unconnected components. Each network will first be introduced and visualized. Later, the networks will be described and compared in more detail. Prominent actors in the networks will be discussed.

**Talking to Networks**

The first riot cleanup network captures only tweets that directly addressed (“talked to”) another Twitter user. It contained 437 ties ranging in strength from 1 to 6, where tie strength is a count of the number of tweets from one member of the pair to the other. It has a large main component, and a number of smaller components. Dyads (components with 2 members) are not shown for visual clarity. For prayer communication, the “talk to” network is substantially smaller. It is quite fragmented, consisting of a number of small components of size 2 to 5, and one slightly larger component. Directly
communicating with others is observed to be the least popular or frequently chosen behavior. Figure 1 shows the riot cleanup network on the left, and the prayer network on the right. Visualizations are created in Netdraw (Borgatti, 2002).

Figure 1. Riot Cleanup and Prayer Networks: Talking To

Quoting Networks

The network of quoted (retweeted) messages was much larger, containing 3035 ties for riot cleanup. Quoting others was also the predominant behavior for prayer communication, though this network is roughly one-third the size of the cleanup network. Both networks have a majority of the actors appearing in the main component. In the visualization, small components are not shown for visual clarity. Figure 2 shows the riot cleanup network on the left, and the prayer network on the right.

Figure 2. Riot Cleanup and Prayer Networks: Quoting
Talking about Networks

This network captures those who mentioned another user in their tweet. For riot cleanup, it was intermediate in size, containing 1526 ties, but had the largest range in tie strength (1-50). For the prayer network, a different pattern emerges. This network is substantially smaller than its companion network, indicating that mentioning others in the context of a prayer hashtag is less common. Figure 3 shows the riot cleanup network on the left, and the prayer network on the right.

Figure 3. Riot Cleanup and Prayer Networks: Talking About

Prominent actors in Riot Cleanup and Prayer networks

For each of these three networks, social network metrics that assess an actor’s centrality or prominence within the network were computed (in-degree, out-degree, in-2step reach, and out-2step reach) using UCINET 6 (Borgatti, Everett, & Freeman, 2002). Degree centrality assesses how prominent, or "central" an actor is, based on how many ties the actor has to others (Freeman, 1979). Indegree captures how many ties come in to an actor from other network members, while outdegree measures how many ties originate with the actor. K-step reach counts the number of nodes that a given actor can reach in k or fewer steps. K=2 finds both neighbors, and "neighbors of neighbors" who can be reached by the actor (Out2Step) or who can reach the actor in 2 hops (In2Step). These twelve metrics were summed to produce an aggregate representation of actor prominence across all three networks.

This aggregate measure was used to select the most central actors in the riot cleanup communications networks from the roughly 4000 actors in the cleanup network, and from the nearly 1350 actors in the prayer network. Figure 4 summarizes the scores for each of the most prominent actors from the riot cleanup communication networks, and Figure 5 does the same for the prayer networks. The most central actors in the riot cleanup communications and in the prayer communications are distinct.

Prominent actors differ in both magnitude and the relative proportion contributed by each social network metric. Actors differentially engage in talking to, talking about or quoting others, and are subject to the same phenomena, being differently perceived by other network members as worth being talked to, talked about or quoted. These directed (Dir) communications, mentions (Men), and retweets (RT) for degree centrality and 2-step reach are weighted equally. None of these behaviors is theoretically asserted as being more critical for establishing prominence in this case.
Patterns indicative of actor roles may thus emerge. For example, an actor with high indegree for retweets and mentions, that is, one who is talked about and quoted frequently, might be a respected member of the riot cleanup community who contributes information and perspectives that are valued by others. An actor with low indegree for these two communicative behaviors, but high outdegree, shares and publicizes other network members and their contributions without being comparably recognized or esteemed by other members of the network. This member's status appears lower. Nevertheless, this network member is still making a contribution to the community.
Figure 5. Actor prominence in Prayer
Actor Similarity

The metrics described previously were used to create an actor profile. Correlating these actor profiles create an actor-by-actor similarity matrix. For riot cleanup, this matrix, rendered as a network (ties indicate actor similarity > .55) in Figure 6, shows the majority of prominent actors are connected in a single component with two main clusters. The remaining actors appear in two isolated dyads.

Figure 6. Actor Similarity Network for Riot Cleanup

Studies of influential Twitter users have observed the dominance of celebrities and mass media in their ranks (Kwak et al., 2010). For this subset of London riot cleanup communications, we find many prominent actors who are neither. In fact, the structure of the large component of the actor similarity network provides an indication of whether an actor might be classified as celebrity/mass media, or whether that actor appears to play a different role.

Of the five actors in the smaller, leftmost cluster of the large component in Figure 7, four - *artistsmakers*, *clean_up_london*, *lawcol888*, and *riotcleanup* – reflect grassroots citizen efforts to self organize to restore the city. One of these actors, *artistsmakers*, has been credited for coining the hashtag *riotcleanup*. The rightmost cluster however contains a number of well-known individuals, including a national level political figure, news media, and entertainers. As there are over 900,000 Twitter users in the data, and nearly 4,000 of them appeared in the network of riot cleanup communications, network information demonstrates value in discovering and clustering interesting actors. Thus behaviorally based social network information and information developed from a derived network computed from actor similarity are used to produce a novel method to identify prominent actors. Interestingly, neither local political figures nor the Metropolitan Police figure prominently in network activity, despite being significant players in managing the riots and dealing with their aftermath. Use of Twitter in this context by law enforcement would not be unprecedented (Heverin & Zach, 2011).

Applying the same method to the *prayer* network highlights differences. While a similarity network can be computed from the most prominent actors, and clusters of similar actors can be identified, there is little further commonality with the *cleanup* network. There are no common actors across these networks, despite both networks being composed of individuals who are prominent is aspects of responding to the London riots. The actor composition also differs for the *prayer* network. The rightmost cluster primarily contains celebrities from pop music or sports, and a few fans. The leftmost cluster includes a number of fans of these celebrities, and lesser-known musicians. No political figures or news personalities appear. No religious figures emerge. Individuals who seem to be oriented toward local action or remediation in the context of the London riots, such as the grassroots activists previously observed are not apparent. While most of the prominent actors in riot *cleanup* are based in or clearly associated with the city of London, this is not the case for *prayer*. The London connection
instead comes from fans and less famous figures. As this study is focused on Twitter activity from the London area, tweets from celebrities who were not in London at the time are not part of the data set. Their out-degree and out-2step reach scores must be zero. Prominent celebrities in prayer thus owe their status to fan response.

*Figure 7. Actor Similarity Network for Prayer*

**Network Characterization**

Several descriptive measures for the six networks under examination are summarized in Table 1 and discussed in the remainder of this section. In some cases, prayer and cleanup for the same communicative behavior (talking to, talking about, and quoting) are compared. In others, overall patterns at the level of the cleanup and prayer networks are considered.

<table>
<thead>
<tr>
<th></th>
<th>Talk to (directed)</th>
<th>Quote (retweet)</th>
<th>Talk about (mention)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cleanup</td>
<td>Pray</td>
<td>Cleanup</td>
</tr>
<tr>
<td>Actors</td>
<td>570 (.15)</td>
<td>128 (.10)</td>
<td>3022 (.78)</td>
</tr>
<tr>
<td>Ties (total)</td>
<td>437 (483)</td>
<td>77 (100)</td>
<td>3035 (3310)</td>
</tr>
<tr>
<td>Maximum tie</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>strength</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main component</td>
<td>190 (.33)</td>
<td>10 (.07)</td>
<td>1963 (.65)</td>
</tr>
<tr>
<td>Maximum indegree</td>
<td>23 (.04)</td>
<td>5 (.04)</td>
<td>115 (.04)</td>
</tr>
<tr>
<td>Maximum outdegree</td>
<td>54 (.09)</td>
<td>5 (.04)</td>
<td>33 (.01)</td>
</tr>
</tbody>
</table>

**Actors.** The number of actors reflects network size. The proportion of actors relative to the total for that communicative behavior is also provided. For example, of all the actors involved in any type of cleanup communicative behavior, 37% were involved in talking about cleanup, but only 15% talked to another actor. In all cases, cleanup networks are larger. The ratio is closest for quoting, and the gap for talking about is largest, with prayer only 14% as large.

**Ties.** Information on both the number of unique ties connecting a pair of actors, and the total count of tweets in the network is provided. Far more ties are observed for cleanup than prayer. The quoting network has the highest number of connections, demonstrating that this behavior is favored.
**Maximum tie strength.** This reflects multiple tweets from one individual to another. This figure is relatively consistent across five of the six networks. The exception is *talking about* regarding *cleanup*, which is roughly an order of magnitude higher than observed in the other 5 networks, showing sustained commitment to sharing commentary on another network member’s actions.

**Main Component.** The main component consists of the largest set of connected actors and their ties. It is the largest component, though it may not contain the majority of actors. The size of the main component and the proportion of actors are reported. The cleanup network shows greater cohesion overall.

**Maximum Indegree and Maximum Outdegree.** In a network, an individual can be more active or prolific in sending ties. Alternately, an actor may be more appealing or focal to others, and receive more ties. Maximum indegree and outdegree identify the highest number of others connected to a single individual. While values for these measures tended to cluster around .04-.06 of the maximum number of potential partners for *cleanup* and *prayer*, there are some interesting exceptions and patterns. Indegree ranges from a low of 5 for *talking to* a single individual about *prayer* to a maximum of 301 *talking about* another person in the *cleanup* context (.04 to .21 of maximum partners). Maximum outdegree has a much smaller range from 5 to 56 (.04 to .09 maximum partners). The relative values of maximum indegree and outdegree for each of the 3 *prayer* networks are quite close (1.0-1.5). None of these networks has a disproportionately focal actor or prolific actor. The relative values of maximum indegree and outdegree for each of the 3 *cleanup* networks show a different pattern. Their respective ratios range from 2.35 to 12.2. Within *cleanup* network for *talking to*, the most prolific actor dominates the ratio. This actor talks to 54 others. It appears that these communications are distributed across actors, rather than being concentrated on a focal, high indegree actor. For *quoting* and *talking about*, the situation is reversed. The dominant actors here are focal actors, receiving ties from 115 and 301 others respectively.

**Discussion**

Notable differences are observed in the social networks for riot cleanup and prayer communications. Riot *cleanup* is far larger than *prayer* in terms of network size and number of tweets. These networks differ in relative frequency and patterns of *talking to*, *quoting*, and *talking about*. They share no highly prominent members. Prominent cleanup network actors included grassroots activists who organized and promoted neighborhood efforts to clear up debris. Here, the network matched real-life roles. For prayer, a mismatch is observed. Prominent actors did not include members of the clergy or other spiritual figures.

Why is the prayer network both smaller and different? Both types of response at first glance are equally reasonable and normal in the aftermath of disaster. A large majority of Londoners are religious, with Christianity the most common religion (Office for National Statistics, 2007). Engaging in prayer requires less energy and effort than cleanup. Since hashtags from both the categories of *cleanup* and *prayer* were among the most common in the data, and were both trending topics, lack of awareness of #pray hashtags does not explain the discrepancy. Other factors must contribute.

A large percentage of riot cleanup communications were positive about riot cleanup and those organizing or participating in these activities, or factual and neutral. Few, if any, were critical of cleanup activity.

```
students' unions and students - if you're near somewhere affected, do your bit and get out and help with #riotcleanup #loveyourneighbourhood

Getting the clean up together - Meet outside Tackle Shop, Roman Road, hackney 9am in the morning to help local shops clean up. #riotcleanup

Today was an amazing experience. Incredibly proud to have been a part of the #riotcleanup and show the world that not all of London is feral

My faith in humanity is restored. Thanks to the good & kind people who get together to clean-up their communities #riotcleanup @riotcleanup
```
The situation for prayer communications was more complex. In some instances, individuals evoked prayer in the context of describing rioting and rioters, and the emotional response they were experiencing.

*scared stiff. Omg the looters are here. My hometown:* (#prayforuk)

*A girl died in a burning house while saving her brother and sister. She sacrificed her life for them. I'm crying.* (#prayforlondon).

Some tweets were clearly sympathetic and supportive of prayer as an appropriate response to events:

*#prayforUK ... my heart with all those affected by the riots - please keep my family in your thoughts*

*Let's all #prayforlondon and continue to heal the world through love, creativity and positivity. Don't let the dark forces win!!!*

However, other tweets expressed a more negative or judgmental attitude towards praying for London as an appropriate response to the crisis. Some found the riots lesser causes for which to invoke a greater power, while others found prayer itself an insufficient response compared to more direct action:

*Vomit! #PrayForLondon is trending! If you're going to hassle God, maybe #PrayForSomalia, #PrayForTalibanWomen, #PrayForChildSoldiers.*

*Don't #PrayForLondon we aren't that lame. We have free education and healthcare. We get paid even if we don't have a job.*

*#controversialmood #londonriots its actually really really boring+ #prayforlondon -blimey how bout actually helpin sort out some root issues*

Some individuals undoubtedly held prior beliefs on these topics, and their tweets may simply reflect pre-existing views. Additionally, actual participation in prayer discussion could be diminished by individual beliefs about the inappropriateness of invoking prayer in the social media context. This would suppress the likelihood of tweeting about prayer as a response to begin with, as well as the likelihood of *talking to, quoting, or talking about* others in that context. Comparable beliefs discouraging taking part in cleanup may not have been present. They are not expressed in the data to any degree.

Social media is inherently not an individual activity. Social influences and forces are likely to have played a role in additional to purely personal views. Through social media, London Twitter users may have been exposed to negative responses to prayer. Individuals who held no beliefs about the prayer response, or held neutral or mildly positive views, would find themselves in a social media environment inconsistent with their position. Burt has discussed how individuals moderate their information sharing and statement of opinions based upon perception of network members’ positions. They may withhold information or bias the information they do share in order to “echo” other’s predispositions (Burt, 2001). Both selective exposure to prayer-related views, rooted in the homophily of social network members, and selective disclosure, driven by a wish to appear consistent with group norms, bias the information environment and influence individual’s behavior in a larger social context (Kitts, 2003). This could discourage adoption of prayer hashtags, and *talking to, quoting, and talking about* behaviors.

Some actors may also need to navigate constraints imposed by formal roles, organizational policy or mandate regarding social media. For example, in the U.S., it was initially unclear if tweeting was a violation of existing Congressional rules on mass communication, and members of Congress remain more likely to use Twitter to broadcast than to engage in dialog with the public (Golbeck, Grimes, & Rogers, 2010). This pattern has been observed for government agencies (Waters & Williams, 2011) and metropolitan police departments (Heverin & Zach, 2011). High-ranking members and spokespersons for these organizations may be sensitive to the implications of evoking and encouraging prayer even in the absence of formal restrictions. Picking up a push broom and sweeping up debris from city streets is less
controversial. Other individuals in the public eye, such as celebrities from the entertainment world and sports figures, may feel freer to speak on a broad range of topics using social media, and may benefit from publicity (Marwick & boyd, 2010).

**Future Directions**

While the presence of a meaningful hashtag is a clear signal of relevance, this method may not capture all relevant tweets. Users might neglect to include the appropriate tag. Use of alternate methods, such as topic models, may help improve tweet recall.

The research examines Twitter communications originating in London beginning with the riots and continuing through the initial weeks of recovery. These communications are used to build networks summarizing the riot cleanup and prayer responses to the crisis as observed in social media. However, the inhabitants of London experienced these events moment by moment. Their emotions, behavior, interpretations, and communications may have changed as events unfolded. Temporal exploration of this data could provide additional insight into this process. Timelines for hashtag categories may show differences. For example, prayer might emerge more quickly in social media data, yet have a shorter half-life and diminish rapidly, perhaps dampened by negative evaluations while cleanup may peak more slowly, and plateau at an elevated rate (Glasgow & Fink, n.d.).

We have treated riot cleanup and prayer as two distinct networks, and observed radically different membership at the top. We have suggested factors at the both individual and social level that might explain this phenomenon. Yet co-occurrence of both types of tags in the same tweet does occur, and these networks do share members. We have not explored frequency or patterns of hashtag co-occurrence or network co-membership. Clustering derived from co-occurrence could produce different sets of tags than the semantically based method that was used, and help refine our understanding of these communities.

Actor profiles are based on a limited set of metrics that capture network information based on actor behavior, response of others to actor, and the actor’s local neighborhood. These profiles could be extended by including additional social network metrics, by adding information on broadcast tweets, or by incorporating connections between network members.

**Conclusion**

This work explores the microblogging landscape of London during the worst period of rioting and public disorder in decades, the riots of August 2011. Through categorizing the hashtags from a corpus of tweets from the city, we identify communications relevant to riot communications in general, and subsets of riot communications relating to specific riot responses, such as cleanup and prayer. Constructing social networks from these tweets based on talking to, quoting, and talking about provides a mechanism for additional insights. Computing social network measures on network actors allows us to identify prominent actors and assess their similarity. We find numerous differences between networks of cleanup and prayer in terms of size, structure and network membership. Patterns differentiating the networks and actors may indicate social processes and forces at work.

For local residents coping with disaster, up-to-date, locally relevant information is critical, and social media may be preferred over mass media (Shklovski et al., 2008). This research shows a more advanced means to purposefully navigate microblogging space than browsing or searching keywords, and provides hooks to engage and share more effectively.

These findings have implications for authorities or emergency responders in terms of improving their overall situational awareness during extreme events. Analyzing relevant social media communications can aid in understanding networks of community communication and response, and in recognizing individuals who play prominent roles in these networks. It can aid in determining if messages are penetrating to the right people, and achieving the desired effect. The role of social media, and ICT more broadly in improving government transparency, and increasing trust and empowerment among the citizenry has been noted (Bertot, Jaeger, & Grimes, 2010) This is an important matter, as lack of transparency regarding the police investigation into the shooting death of a Tottenham resident motivated the protest that sparked the riots.
References


Granovetter, M. S. (1973). The Strength of Weak Ties. The American Journal of Sociology, 78(6), 1360–1380.


Office for National Statistics. (2007, January 1). Table View. Retrieved May 12, 2012, from http://www.neighbourhood.statistics.gov.uk/dissemination/LeadTableView.do;jsessionid=d9W9PnnNcyZ11Yw2J1bYWDPHqJ2wNLpZz6mNLnKM0JsZpxqySJY-11496570761336862605420?a=7&b=6059498&c=london&d=81&e=13&g=336453&i=1001x1003x1006&k=religion&o=254&m=0&r=1&s=1336862605420&enc=1&domainId=16&dsFamilyId=95&nss=true&nsck=true&nssvg=false&nswid=1600


Collaboration in Isolation: Bridging Social and Geographical Boundaries in Two Rural Technology Firms

Sean P. Goggins
Drexel University
outdoors@acm.org

Abstract

Little attention has been given to understanding how technology is used to bridge social and geographical boundaries between metropolitan and rural firms in developed nations like the United States. In this paper we draw on theory from social worlds to contrast the introduction and integration of collaboration technologies in two small (23-53 employees) rural technology firms in the Midwestern US. We report on a three-year ethnographic study of the work practices and technologies used to forge relationships between distant offices and customers at each firm. Data include participant observation, field notes, memos and interviews with 71 informants. Based on our data, we identified organizational facets of geography, socio-technical configuration and collaboration practices as instrumental in the formation of working spheres \[31\] for members of each firm. As the facets of each firm change over time, our data show social, organizational and technical practices lead to different types of virtual organization at each firm. One firm’s working spheres coalesce, while the other firm’s working spheres split. We show that small firms operating in rural areas experience unique social and geographic boundaries in their development as virtual organizations.

Keywords: rural technology, distributed work, social worlds, CSCW

Introduction

Small technology firms in rural regions are developing new forms of virtual organization by using information and communication technologies (ICTs) to collaborate with metropolitan partners and customers. These firms bridge long standing social and geographic boundaries between metropolitan and rural areas (Williams, 1973). Physical distance is the most defining characteristic, but the social distance between rural society and urban society is an ever present and under examined phenomena (Malecki, 2009; Williams, 1973), that the socio-technical frame of information science is uniquely positioned to address.

There are many case studies documenting the first mechanism of virtual community (Bers, 2001; Blanchard & Markus, 2002; Ellis, Oldridge, & Vasconcelos, 2004; Gabriel, 2004; Healy, White, Eshghi, Reeves, & Light, 2007; Rohde, Reinecke, Pape, & Janneck, 2004), and an equal number that shed light on the collaboration practices of nomadic and geographically isolated individuals who connect to an organization from afar (Gloor, 2005; Gutwin, Penner, & Schneider, 2004; Mark & Su, 2010; Scacchi, 2007; Turner, Bowker, Gasser, & Zacklad, 2006). A small number of studies focus on the influence of ICTs on economic and social development at the scale of a nation or region in the undeveloped world (Ali & Bailur, 2007; Friedman, Kahn Jr, & Borning, 2006; Mark, Al-Ani, & Semaan, 2009). Literature in MIS is replete with studies of large firms outsourcing operations. The use of ICTs to overcome or at least mitigate distance is a theme that runs through each research community (Olson & Olson, 2000).

However, there are no prior comparative socio-technical case studies of the uptake and use of ICTs by small firms located in rural areas.

Acknowledgement: This work was supported in part by NSF Award number 1221254, Toward a Context Adaptive Theory of the Relationship Between Structural Fluidity and Virtual Organization Performance, from the NSF Office of Cyberinfrastructure, Virtual Organizations Program. We would also like to thank our anonymous reviewers and the journal editor for their feedback, which substantially advanced this article. We specifically thank Mick Khoo and Christopher Mascaro for their comments on earlier versions of this article.


Copyright is held by the author.
Small, isolated rural firms are an important type of case for socio-technical researchers because these firms, out of necessity, have to construct complex working spheres to span these social and geographical boundaries. A working sphere is composed of a collocated team, a distributed team and an organizational structure (2004). The two firms in this study consist of collocated teams located in rural regions of the US, whose customers are distributed teams that are located in both metropolitan and rural areas. How the working spheres in these firms use ICTs to overcome traditional social and geographical boundaries stretches our understanding of virtual organization and the role of geography. More widely, the economic development plans of scores of communities around the world (Maitland & van Gorp, 2009) will benefit from examples of firms connecting metropolitan and rural areas. Understanding how these firms overcome traditional barriers is therefore of immediate value to both citizens and scholars. In this paper we analyze two ethnographic case studies of rural technology firms in the Midwestern United States. We also describe how understanding distributed work practices in firms with different geographical and economic traits might inform our understanding of virtual organizations more generally.

Related Work

Defining Rural

In this section, we define "rural" more precisely, and contrast the distinct views of technology adoption found in regional studies and organization science literature. Gilligan (2005) critically examines rural adoption and use of ICTs by exploring what we mean when we say "rural", and how rural sociologists define "rural" as at least partly a social construction. “Rural” is a category of thought, viewed as unwavering against the forces of change and time, while “urban” is a category of thought which marches forward and defines the next age. These categories influence perception of a region. Gilligan (2005) found that ICT uptake and use varies greatly in rural areas, and greatly according to the specific technology; but perception of uptake is uniformly lower for rural areas, even in cases where uptake is rapid.

To tease out differences between specific rural areas, regional studies literature focuses on ways to measure and classify degrees of “rural”. In the United States, where the two firms we examine are located, a standard measure of rurality is metropolitan influence, which the United States Department of Agriculture (USDA) prescribes as a continuum of 1 (most metropolitan influence) to 12. Comparing Gilligan’s social construction of rural on this continuum, we see “1” reflects an “urban” social world, and 2-12 reflect increasing levels of social “ruralness”. When we talk about rural, we are referring to these classes, 2-12.

Organization, Region and ICT

With few exceptions (Tapia, 2004), prior studies of collaborative computing tool adoption take metropolitan infrastructures, social configurations and lifestyles as a given. The organization science literature includes analysis of how ICTs are taken up in large, complex organizations (Mark & Poltrock, 2004; Orlikowski & Barley, 2001; Finholt & Sproull, 1990), how this use reframes and restructures power and roles (Barley, 1986; Nardi, Whittaker, & Schwarz, 2002; Eschenfelder, Heckman, & Sawyer, 1998; Lamb, King, & Kling, 2003), and how work groups are constituted in technologically mediated ways (Brown & Duguid, 1991; Keisler & Cummings, 2002; Olson & Olson, 2000; Muller & Gruen, 2005).

Regional studies literature often restricts its analysis to observations outside the boundaries of specific firms. Measurements are chiefly macro-economic, and the internal characteristics and practices of different firms are unexamined. In regional studies research, technology appears in terms of the “digital economy” and the “analysis of ICT diffusion into isolated communities” (Hollifield & Donnermeyer, 2003). For instance, Malecki & Moriset (Malecki & Moriset, 2008, p. 200) define the digital economy as the pervasive use of ICTs. Moriset argues that networked connections to metropolitan areas, combined with widespread availability of ICTs make rural technology firms likely to succeed (2003).

In contrast to Moriset’s optimism about the future of rural technology firms, there is evidence that pioneers in this space struggle (Malecki & Moriset, 2008, p. 200). Rural firms must overcome the social
and distance barriers described above. Some concern exists about an urban-rural “digital divide”, but this is a home access gap, not a business access gap (Horrigan, 2009). Engaging the digital economy requires rural areas to overcome the intrinsic penalty that arises from their isolation, and perception of scarce human capital is the most limiting factor (Malecki & Moriset, 2008, p. 11). The rural human resource gap includes genuine labor scarcity and perceived scarcity, depending on the region in question (Henderson & Abraham, 2004).

Technology firms that thrive in rural communities have to overcome social, geographical and resource barriers, but how they do so is unexplored. A reflexive study of technology uptake and use by rural technology firms will fill important gaps in two research communities where collaborative computing plays a pivotal role. Such firms grow quickly, are dynamic, complex and highly distributed; the embodiment of virtual organization and rural economic impact. Malecki (2008) characterizes firms who locate in smaller economies and bring money in from metropolitan areas as sophisticated consumers of space.

Previous Inquiries of Distributed Work Teams

Rural work teams interact with groups in their local office, customer groups, and groups at other offices. Faced with distributed work inquiries similar to ours, past researchers have used social worlds theory, communities of practice and ensembles, a proposed additional analytical level in Activity theory. In this section, we briefly review the application of each of these theories in prior studies of distributed work teams.

The development of virtual workgroups within the same organization vis a vis ICTs is explored by Mark and Poltrock (2004), who examine instant messenger adoption across locations within a single organization. Their study focuses on firms recently merged into Boeing using the construct of social worlds (Denzin, 1978). They use Clarke’s (1991) definition of social worlds: groups with shared commitments to certain activities, sharing resources of many kinds to achieve their goals, and building shared ideologies about how to go about their business. Each member’s workgroups are framed as distinct working spheres, which are a special type of social world that focuses on the work of the organization.

![Figure 1 - Mark & Poltrock's (2004) working spheres](image)

Explaining their construct of working spheres as an extension of social worlds, Mark and Poltrock (2004, p. 305) point out that it frames the study of distributed, highly virtual organizations without the premise of the intricate and longitudinal mutual engagement required by Communities of Practice (Wenger, 1998). Further, working spheres incorporate the local characteristics of each group more directly than social capital theory (Coleman, 1988) or social network analysis (Granovetter, 1985; Wasserman, 1994).

Technology can also take a more central role in adoption studies. For example, activity theory creates space for the role of technologies and artifacts in distributed work and is used to frame CSCW and HCI research. Many researchers struggle with dynamic group work and whether to classify aspects of it as either action or activity. There are limitations to how action or activity may "account for the way individuals in practice conceptualize, delimit, and represent those practical intermediate units of work that
allow them to instantiate their activities through sets of thematically connected actions” (2009). In an effort to apply activity theory to more dynamic group constructions where work efforts encompass actions but are not as clearly object related as activities, Gonzalez, Nardi & Mark (2009) explicate the construct of ensembles. In this conceptualization of activity theory, an ensemble is an “intermediate unit of work between actions and activity in the hierarchical framework proposed by classical activity theory” (Kaptelinin & Nardi, 2006).

Ensembles are sets of thematically related actions defined by a purpose (Gonzalez et al., 2009, p. 110), and in this way are more adaptable than standalone concepts like working spheres, projects and tasks. Ensembles help to make actions meaningful beyond their immediate scope, and serve to make workloads visible within teams -- both powerful explanatory attributes in the study of virtual teams. When groups lack intimate, longitudinal relations, as many of those examined here do, working spheres is more salient.

**Literature Summary**

Working spheres and ensembles are constructs that contribute a theoretical frame to the study of collaboration in rural technology firms. Each is applied in established, structured organizations and demonstrates the explanatory power of good theory. This study is informed by these theories, but takes a reflexive stance toward the data.

The rest of this paper is structured as follows. In the next section, we describe the field setting and research methods in greater detail. Following that, we compare and contrast the two firms from three perspectives: geography, socio-technical structure and distributed work practices. Throughout that section we describe how distinct forms of virtual organization develop in each case, and how these developments are influenced by degrees of isolation and business model. Implications for research and practice are presented at the end.

**Case Descriptions and Methods**

**Rural Technology Firms**

In this paper we examine two cases of virtual organizations that emerge around two firms in rural, Midwestern US communities. The first organization, Medium-City-Co (MCC), is in a small city, and the second organization, Small-Town-Co (STC), is in a small rural town. From 2006 to 2009, we conducted ethnographic research at STC, and extended this study to daily observation of the work at STC for 8 months in 2009. The two firms we select generate at least 70% of their revenue from outside of their immediate area, indicating significant ongoing connection with metropolitan areas (1996).

MCC and STC are smaller than the organizations Mark and Poltrock (2004) examined and have different pressures, but they do share geographic isolation as a defining trait. MCC and STC experience social and distance barriers between nodes in a virtual organization, but also as a more absolute isolation from modern, metropolitan economies.

**Field Setting: Situating the Cases**

The nature of each firm’s location is material to each case internally, and integral to the comparisons we draw between the cases. MCC’s main office is located in a small city with a population ~90,000, located 100 miles from the nearest city with over 150,000 people. STC is in a small town with a population ~4,000, located over 200 miles from the nearest city with over 150,000 people. Our data collection and analysis incorporated all locations for each firm, as summarized in table one.
MCC was founded in the early 1990’s in a small, Midwestern college city of less than 100,000 people, more than two hours from the nearest metropolitan area. Their principle, current business is a hosted, educational software product. Between 23 & 25 people worked at MCC during the study period. STC was founded in 2004 in a small Midwestern US town of less than 5,000 people, more than three hours from the nearest metropolitan area. STC was conceived from the beginning as a company that would move technology jobs from large metropolitan areas to rural communities; its revenue is from IT outsourcing. Employment at STC has ranged from 8 to 53 employees during the period of our study. STC’s main business is IT outsourcing. Management and organizational structure at both firms is centered on the single entrepreneur who started the company. MCC’s proprietor is Horace, and STC’s proprietor is Alan. Each location has an employee responsible for administrative tasks, but most work at both firms is organized in small, self-directed groups of 3-5.

Methods

The principle data collection method was ethnography. Organizational ethnography, particularly if carried out with a significant longitudinal dimension, is capable of generating richly textured data on organizational members’ daily working practices, leading to insights into wider organizational culture and behaviors, such as organizational responses to changes in technological environments (Barley, 1986). From 2006 to 2009, we conducted ethnographic research at STC, and extended this study to daily observation of the work at STC for 8 months in 2009. Between 2005 and 2009, we conducted ethnographic research using participant observation, interviews, and documentary analysis to build a rich understanding of the social, working and information spheres of these two technology firms. From 2005 to 2008 we participated several days a week for 40 months in daily operations, meetings and organizational functions at MCC. In both cases we conducted and recorded interviews, observed meetings, and maintained field notes and memos. For data analysis, we followed LeCompte & Schensul’s (1999) guidance for the interpretation of our ethnographic data. In the field, our data analysis focused first on description of how each organization worked and how members made sense of the socio-technically constructed relationships with distant partners, customers and locations.

At both sites our “in the field analysis” integrated a constant awareness of the “hermeneutical process” principle. The hermeneutical process foregrounds a constant consideration of the interdependent meaning developed through interaction with and interpretation of the environment. This process is central to our analysis and the experience of the organizations we studied. Out of the field, data was analyzed first from the bottom up using a grounded theory approach. The socio-technical facets we identify later in the paper emerge from this coding. The data was then analyzed from top down using a coding scheme for organizational traits, identified initially during our bottom up coding. The facets of geographical isolation emerge from these codes. Constant comparative analysis was used to surface the socio-technical facets and unique work practices that distinguish the two cases.

Interpreting and contrasting these case studies from a socio-technical practice perspective involved drawing on three methodological approaches. First, grounding our understanding of what each

Table 1 - Study location summary (December, 2008 Census)

<table>
<thead>
<tr>
<th>Location</th>
<th>Rurality¹</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium-City-Co HQ</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>Medium-City-Co Satellite 1</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Medium-City-Co Satellite 2</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Small-Town-Co HQ</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>Small-Town-Co Satellite 1</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Small-Town-Co Satellite 2</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>

¹ USDA categories, numbered 2-12 for rural areas
organization member did at work was accomplished through extended, daily participation as a member of each organization, in the ethnographic research tradition. Second, accounting for relationships with local organizations, customers and distant offices was accomplished through participation across this range of the organization. Third, through regular analysis of our data in the field we ensured that changes in socio-technical practices over time were recorded, noted and surfaced in this analysis.

Results and Analysis

This comparative case study shows how social and geographic barriers are bridged and circumvented by small, rural firms. Based on our data, we identified organizational facets of geography, socio-technical configuration and collaboration practices as instrumental in the formation of working spheres (2004) for members of each firm. We describe these in the sections that follow. As the facets of each firm change over time, our data show social, organizational and technical practices lead to different types of virtual organization at each firm. MCC’s working spheres coalesce, while STC’s working spheres split. This is woven into the three sections below, and illustrated in figure 2.

![Figure 2 - Medium-City-Co (left) and Small-Town-Co (right) high level Working Spheres differences](image)

Facets of Geography

The grounded theoretical coding and analysis of the field data (Glaser & Strass, 1967) led to the identification of the core category of ‘geographical isolation’ as being central (in both positive and negative ways) to the behavior of both firms. Further analysis and coding of this core category led to the identification of 3 subsidiary (or ‘axial’) dimensions, which were found to be important for each firm.

Physical Workspace

Both firms benefited from low cost office space, compared to large metropolitan areas. In the case of MCC, between three and five employees operated out of basement headquarters during the first four years of operation. In the fifth year the firm moved into leased commercial space where it remained through the period of our study. STC’s startup operation required no physical space initially, and within six months the firm was working out of commercial office space. MCC’s physical space costs were 30% of what they would have been in a large metropolitan area; STC’s were 25%. These fixed costs of growth in more isolated regions were lower, and both proprietors indicated that this influenced their perception of risk, compared with comparable metropolitan operations. Lower costs were a positive feature of isolation.

Worker Mobility

Worker mobility in small cities and small towns takes on a uniquely protective flavor in the two cases. While “no competition agreements” are common in engineering and technology professions in the
USA, the specific cost damage clauses for these two firms is punitive, suggesting a perceived threat not simply of the loss of an employees utility, but the potential generation of a new and particularly dangerous local, competitive threat. For example, MCC’s agreement calls for a $1,000/day penalty for each day a former employee works for a customer or competitor for two years after employment ends.

Unlike metropolitan firms, where the primary threat of worker mobility is employment with a rival or customer, the threat perceived by technology firm proprietors in isolated areas is more akin to the threat of property loss. For both firms, the no-compete agreements restrict employment within a 50-100 mile radius of the organizations offices. The enforceability of such agreements may be debated in other forums. The act of reading and signing the agreement does, from interviews with both firms’ employees, create a clear impression of the boundaries of post-employment behavior in these isolated regions, and influences decisions about when and under what circumstances employees choose to terminate their employment.

The Vital Network

A third distinguishing characteristic in the experience of isolated technology firms, in contrast with their more metropolitan counterparts, is that MCC and STC rely on technological network connections to the world beyond their region to perform work. This shapes their working spheres in two distinct ways. First, at both firms operations literally cease if there is an interruption in the network connection between the firms and their geographically dispersed customers. Second, the significant relationships that develop with customers and across firm sites are fundamentally socio-technical. Teams work together at the same time, in different places, almost exclusively through ICTs.

Socio-Technical Configuration Facets

In this next section, we describe the combination of internal practices and technologies that these firms developed in response to this isolation. Each firm developed its external relationships in a way that reflected the intersection of the firm’s isolation, business model and available technology. We describe these distinctions as ‘socio-technical facets.’ While the external socio-technical facets are distinct, the two firms share many socio-technical facets corresponding with internal, interoffice communication, coordination and information sharing. However, their communication practices vary widely, and it is the external forces that drive the key differences in these latter socio-technical facets. First, therefore, we examine and discuss the primary, common internal socio-technical facets. Second, we discuss how customer needs drive differences in the customer facing socio-technical facets for each organization.

Common Internal Socio-Technical Facets

As distinct as the socio-technical facets are outside of their locale and boundaries, these two isolated technology firms built their interoffice connections with a common arrangement of technologies. For synchronous communication, each firm uses a combination of Skype and Microsoft Instant Messenger. Instant messenger is used for interruptions between employees and Skype is used for planned interoffice communication. To coordinate tasks, both firms relied on a common set of tools as well. During project initiation, a hosted software product called Basecamp is used to develop task lists and assignments. Once a project is launched, both firms use a tool called FogBugz to manage requests, assign work and communicate with internal technology workers and customers. Not surprisingly, these common technology choices, which emerge from Internet searches for software management tools, are adopted in different ways and play different roles in each firm.

External Customer Differences Drive Faceting

One socio-technical factor that shaped the adoption of information technologies (particularly groupware) in different ways in each firm was the relationship between each firm and their customers. In this section, we start to see how the use of ICTs in work practice was both shaped and shaped by the
working spheres of each firm in different ways. These socio-technical facets are a significant influence on the way each firm constructs a virtual organization that incorporates the firm and its customers.

For MCC, the key socio-technical facets are “emotionally safe communication”, “scripting interactions with customers” and a stable set of collaboration technologies. MCC provides a service that the customers rely on, but do not understand; a subscription based, hosted software product that now serves over 30 colleges and universities. Strategic use of ICTs enabled this line of business to grow without requiring travel to client sites for sales, installation or support. For most of these 30+ software-as-a-service customers, the software configuration, management and support are handled entirely using ICTs. This requires MCC’s employees to incorporate technology in a way that provided a predictable experience for customers.

Security policies, time accountability and practices of interruption management are the principal socio-technical facets at STC. In contrast to MCC, STC’s need for information and communication technology to reach beyond their locale is core to their opening business model. STC has five major customers who are responsible for more than 90% of their gross revenue. Each of the major customers has a different set of collaboration tools required for work within those customer sites. Consequently, STC’s employees routinely adapt to different socio-technical infrastructure facets, depending on which customer they are interacting with, which influences usernames, passwords, time scheduling and how they are interrupted in the local organization, compared with the extended, virtual organization that incorporates the client.

These two firms have significant differences in their socio-technical faceting, though they use similar technologies. MCC is the collaborative technology leader who is able to define practices within itself and for its customers. STC is a collaborative technology consumer that adapts to unique socio-technical configurations on a customer-by-customer basis. Though each firm leverages the opportunity of lower cost structures and a less mobile workforce combined with high speed networking access into metropolitan areas, the way these factors impact the experience of the firms is influenced by how they structure their social interactions, and how ICTs are used to mitigate distance.

The different socio-technical facets that distinguish these two firms arises from and reinforces differences in communication flow between the firms and their customers, depicted in figure 3. A MCC employee uses a consistent and stable set of information and communication technologies, and those technologies are core to the collaboration and communication that occurs between MCC customers and MCC. STC employees reference a shifting and varied range of ICTs both internally and externally, depending on the context of the work.

Figure 3 - Communication Technology Use at Medium-City-Co and Small-Town-Co
Collaboration Practice Organizational Facets

In each firm, small work groups interact with each other and with customer groups outside the firm. The practices developed through these interactions create dynamic and sustained working spheres. The socio-technical facets for each firm emerge from the structuring influence of their technology configuration, and from the collaboration practices that emerge.

Medium-City-Co

MCC’s collaboration emerges around two major functions of their business: Software design and sales. The owner, Horace, is involved in both, as is Bryan, who manages a good deal of customer contact. In the next sections we describe the facets of collaboration in these two groups, and through this description MCC’s socio-technical facets of emotionally safe communication, scripted interactions and stable technologies become clearer.

Software Design Groups at Medium-City-Co

MCC regularly engages customers in design discussions and decisions to improve their product. Bryan, who is the primary customer liaison, operates as a proxy for customer software change requests to the Schoolsoft (not the real name) hosted application that is now MCC’s dominant source of revenue. Horace, the owner of the company, works closely with Bryan during early design processes, usually following Bryan’s conversations with customers. Quinton, the chief architect for the product, is an accomplished technologist, and usually participates in design sessions. Design sessions are planned, as Quinton and Bryan are at the opposite ends of 6,000 square foot office in the Midwest, and Horace does his work from another state.

In this next section, we describe an example of the design process at MCC. This micro-case begins with Bryan and a customer with 400 active users trying to build an organization level report. Bryan and the customer do this using WebEx, a telephone and a digital sketchpad. Schoolsoft is designed to support individual user needs, but not the assembly of data pulled together from multiple users. Bryan and the customer each expect this limitation can be overcome. In this case, Bryan sketches a basic system design (user interface) that, in concept, enables the customer organization to manage member data and produce reports at the organizational level.

Bryan’s sketch could be viewed as a boundary object between MCC and the customer, except that Bryan’s use of the sketch reflects simple articulation of a shared understanding. Bryan and the customer already share much common ground, the Schoolsoft system, and the sketch is within that common ground. The boundary for Schoolsoft customers is the emotional tension of dealing with a technology product, and a technology company. Sketching is a recognized interaction design technique, but it is also viewed by customers engaged in the design process as non-threatening, or in terms of socio-technical faceting, emotionally safe. This is one way MCC bridges the social barriers between themselves and their customers.

With the ideas of the customer clear, Bryan, Horace and Quinton begin a design session. One notable characteristic of these design sessions is the fluid role switching that occurs. Throughout each design interaction, there is “sketcher”, a “problem space owner” and an “implementation analyst”. Bryan is most frequently the sketcher, Horace most frequently takes on the role of the problem space owner, and Quinton is most often the implementation analyst.

During 17 of the 23 design sessions we observed, there are periods where the members jump into different roles, and when this happens another member fills the missing role. The members rotate, in a sense. This occurs most fluidly during distributed design sessions, when Horace is at his office in a different state; 19 of the design sessions take place with Horace at a remote office. During one session, Quinton prepared a sketch of an idea he developed to implement an improved “customer dashboard”. Horace immediately jumped into the space of implementation analyst and Bryan rejoined the role of problem space owner. They proceeded in these roles for 80 minutes, leading to a design that was eventually implemented in the Schoolsoft system. While these transitions also occurred during sessions when the team was physically co-located, the dynamic shift of roles was frequent and clear only during
the distributed collaboration work. In this way, socio-technically executed design practices at MCC appear to enable this group to work more effectively in a distributed manner than a face-to-face manner.

**Customer Groups at Medium-City-Co**

As noted in our discussion of the software design team, the transitions between roles and production of designs flowed more smoothly when the team worked in a distributed manner. In the case of customer development (sales), a physically co-located team in the Midwest handled this repetitive work, which consisted mainly of methodically contacting, following up and providing demonstrations of the software product for prospective customers. When Horace and others at his remote office became involved in these established sales processes a new type of coordination challenge emerged. Despite numerous attempts at resolution, challenges and ongoing communication failures persisted whenever the sales process became distributed across locations.

At first blush, this confirms well-established findings in CSCW: distance matters. What is novel here is where it matters. Software design is an unstructured activity. At MCC the sales process is highly structured and repetitive. Horace & Bryan are active in both processes. The same tools are used for both processes. The barriers of distance are constant. Distributed design works well, distributed sales work does not. How is this possible? The principle difference is how the working spheres are constructed in the sales process. With sales there are more working spheres, and they change more frequently. In order to overcome distance and social barriers between the firm and customers, MCC builds an effective distributed design enterprise; they do this because they have to. With the sales process, they simply keep it in the main office because there is no penalty. The lesson is that barriers fall when survival depends on them falling, and MCC’s distributed design success exemplifies that.

**Small-Town-Co Collaboration**

STC work groups emerge around the two key practice domains of sales and customer projects. Each of these functional teams is constructed around a stable core of STC management personnel: Alan, Bob, Carl, David, Ellen, Fred, Gerritt & Helen. Each functionality oriented group experiences the socio-technical nature of distributed collaboration differently. Unlike MCC, each STC employee must adapt and make use of different technologies for each customer interaction, including in some cases customer specific email infrastructure. In the next section we describe the practices and trajectories of socio-technical group emergence at STC in the two key practice domains where we observed extensive group work. Through these descriptions, the dominant socio-technical facets of customer security policies, time accountability and interruption practices emerge.

**Sales Process Groups at Small-Town-Co**

STC services are sold across a wide array of industries and geographic locations. While MCC leverages information and communication technology to facilitate sales across the US, STC’s markets expect face to face visits, which in turn requires an average of 2-3 days of sales travel each week for Alan, the owner. Unlike MCC, sales practices are not reified in a sustainable way.

Isolation and being in a rural area is not handled with ICTs at STC; people travel. This condition reverberates through each of the working spheres at STC. Practices for assessing and sharing information or coordinating sales activities are spread across the organization and not handled consistently. For example, practices for assessing information quality in the prospecting cycle, selecting new cities to target and managing the process of moving a sales lead to close are all highly variable. Three factors contribute to these performance gaps. First, heavy travel to meet customers and potential customers limits the amount of time used for planning. Second, like many technology firms, STC’s internal IT is not well developed. Third, operations management and senior technologists are often pulled into the sales process with limited preparation. This third point has repercussions in the broader organization. Operations staff members (people on teams doing customer work) are frequently interrupted, which interferes with performance on customer contracts and creates a tension between
customer requirements for time logging and internal billing – 40 hours per week – and non-billable, unplanned sales work.

Customer Project Teams at Small-Town-Co

Following a successful sale, work moves into project execution. The main type of deal is their preferred “outsourcing” deal that includes the services of a minimum of four personnel for at least three months, but usually for two years or more. Project types include ongoing software maintenance, software testing or data analysis services for the customer. The transition between sales and operations follows a set of established practices. Bob, Carl, Gerritt and Fred are all experienced at successfully establishing customer specific socio-technical infrastructure, which typically involves the configuration of new software and the establishment of new coordination practices with a customer. Unlike MCC, STC does not dictate communication and information sharing technologies; each team adapts to its customer.

This adaptation to customer technology includes STC employees becoming primarily accountable for time directly through the customer system, use of customer security infrastructure and participation in social technologies like email and instant messenger exclusively through customer remote desktops. The consequence for STC employees is that their identity shifts quickly to that of the customer; their connection to STC diminishes. This results in coordination challenges within STC, and across teams serving different customers. Here we see an inversion of the usual consultant experience, where a nomad loses touch with his employer because he is never there. In this case, people sitting side by side in the same room lose touch with their employer because they are connecting to an outside world through an ICT. The person seated next to an STC worker is less connected than the person on IM, in a large city.

Outsourcing customer teams at STC do not experience challenges with awareness and interruption management resulting from remote work; at STC such challenges are local. The STC workers and their customers develop a repertoire of coordination and information sharing practices that serve the customer work well. The challenge with interruption comes from the co-located management and other work teams, who rely on STC workers for sales support and internal operation support regularly. The close identity workers develop with the customer they are working for is placed in regular conflict with local demands. STC workers often work extended shifts to manage these demands because their personal, geographic isolation means that STC is their only choice for a technology job.

Discussion

The question of how groupware technologies diffuse through and support work groups in organizations is an important one that has received a lot of attention. As the socio-technical landscapes within which such diffusion takes place are constantly changing and evolving, there is always a need for new studies of these processes. The ready availability of cheap collaboration tools enabled MCC and STC to create unique forms of virtual organization in isolated regions. Each followed a unique trajectory of adoption and use.

Mark and Poltrock provide one useful analytical lens for analysis of the diffusion of groupware in organizations. They correctly point out that many existing theoretical approaches to the study of organizational groups reify the concept of the group itself, to the point where the fluid nature of such groups is ignored. They propose the concept of the working sphere, a fluid, emergent, ad hoc and shifting model of the temporary alliances that form and dissolve in response to ongoing organizational tasks. It is the circulation of organizational members across working spheres that provides a path for individuals (and the groupware that they adopt) to travel and diffuse through an organization.

This model is useful but it also has a number of limitations, and these provide opportunities for new research. The original analysis focused on Boeing (2009 revenues, $68 billion) and bracketed off a large number of socio-technical factors at this site as something of a given, such as the large scale of the firm, its metropolitan locations, etc. What therefore is to be learned from the application of working spheres in other contexts, such as the small-scale, geographically isolated firms? Our contrast begins with the geographical isolation of these two firms, and extends to the work practices that develop in the distinct working spheres that emerge within and around these firms. 
In these two, geographically isolated firms, technology is not simply adopted, it transforms what is possible. For instance, MCC’s decision to supply turnkey software services led them to deal with a standard product, which led to the routinization of a lot of their business and organizational processes, and thus also to the routinization of their groupware tools and use. This was reflected, for instance, in the ways in which team members fluidly switch roles during creative work. In contrast, STC’s business model of providing outsourced IT support constantly placed their staff in new organizational and technological environments, in which their clients controlled a number of important environmental factors. This meant that STC had to adapt constantly to new situations. While this ‘agility’ might be seen as advantageous (and perhaps emblematic of new types of firm in knowledge and service economies), it was also problematic, in that this ongoing external shaping effectively prevented the routinization of many of their core organizational practices.

In the case of MCC, therefore, situational factors led to organizational consolidation, routine, and stability; in effect, they approached the same problem time and time again. In the case of STC, on the other hand, situational factors led to change, churn, and a lack of stability and routine; there was always something new to deal with. We can view this contrast between the two firms in terms of an adaption of Mark and Poltrock’s original figure (Figure 1, above). In this adaptation, we can see that in the case of MCC, there were centripetal organizational forces – such as the ongoing routinization of organizational process in relation to a standard product – that acted to pull the work groups and their associated groupware back towards the core business, and thus also back together. In the case of STC, there were centrifugal external organizational forces – such as the constantly changing technological practices of their clients – that acted to pull them apart.

We represent this conceptually in figure 4. As in Mark and Poltrock, the circles are working spheres, populated with group members. We have added two new elements: stars, representing environmental factors that either pull groups together, or pull them apart, and arrows, representing the pulling forces. The degree of overlap in the spheres defines the structure of the virtual organizations that form around each firm. MCC is represented as having developed a set of socio-technical facets and work practices that draws remote customers into their way of working, leading to the routinization of organizational practices across the firm, represented by the star. It can be seen that the work spheres overlap considerably. At STC, in contrast, members are pulled toward external customer concerns (represented by multiple stars), and as a result they participate in working spheres that often barely overlap, or which may even be in conflict.

The varying degree of overlap in working spheres has a number of consequences. First, the greater geographic isolation experienced by STC workers creates greater contention and accountability for their time with customers, but also leads to more local interruption. MCC bridges the social divide by creating routine. STC bridges the social divide by frequent face-to-face visits with distant customers. Geographical and social barriers are overcome in each firm by establishing working spheres that incorporate local and distant members. These firms do not adopt ICTs and working spheres do not resist; instead, two different classes of rural technology firm emerge through their use. ICTs in these firms, and
we suspect other isolated technology firms, are adopted quickly and transform what is possible at a speed that few studies of large corporations have noted. We think it is possible that these kinds of firms could become important organizational laboratories for CSCW research in the next decade.

**Structural Fluidity in Rural Technology Firms**

Working spheres are a theoretical construct that enable reflection about how new forms of structurally fluid organization are enacted through technology and across geographic boundaries. Organizational studies have examined dynamic work groups, organizations and distributed work. Distance between work groups and the efficacy of technologies for bridging that distance are the subject of ongoing inquiry. In the case of these rural technology firms, we show two cases where structural fluidity – the capacity of an organization to shape itself, through technology, into a context that includes customer organizations – supports new organizational shapes in new types of regions; or, at least regions that are under examined in the organization science literature. Developing theory to better understand how structural fluidity is connected to performance in a range of organizational contexts in the future.

**Conclusion**

The distributed work in the two small, geographically isolated firms described in this paper takes a notably different form than that which has been described in prior studies. The differences in their socio-technical practices, and the rapid rate at which they adopt new technologies to structure their businesses, suggest that other small geographically isolated firms that operate as virtual organizations may also take multiple complex forms. The contrast between these two firms and previously reported examples suggests that the uptake and use of information and communication technologies now enables a new class of virtual organization that warrants further study.

**References**


Group Polarization in Virtual Communities: 
The Case of Stock Message Boards

Hsuanwei Michelle Chen
School of Library and Information Science, San Jose State University
hsuanwei.chen@sjsu.edu

Abstract
This study examines group polarization, the phenomenon that group members tend to become more extreme in their decisions after group discussions, in the setting of virtual communities, specifically on stock message boards. In this paper, I propose two factors – group size and discussion thread length – to introduce the group polarization theories, social comparison theory (SCT) and persuasive arguments theory (PAT), respectively, to virtual communities. I examined the effects that group size and thread length have on group polarization, which is measured by forum participants’ sentiments. The findings suggest that group size has a significant effect on group polarization. The result of thread length, however, is mixed: the effect of thread length depends on the degree of posting activity on the message board. The analysis and findings echo prior literature, in which SCT and PAT are claimed to serve as two key mechanisms that drive group polarization.

Keywords: virtual communities, group polarization, social comparison, persuasive arguments

Introduction
There has been a long interest in studying the phenomenon of group polarization in psychology and social science literature. Group polarization refers to the phenomenon that group members have a tendency to become more extreme in their thinking (positions, decisions, or choices) after group discussions (Isenberg 1986; Myers and Lamm 1976). For instance, group polarization has been widely adopted to explain the jury decision-making process: jury members usually decided on punitive damage awards that were larger or smaller than the amount any individual juror had preferred prior to deliberation after group discussions. Group polarization, therefore, presents the effects of extreme inclinations of people after deliberating with each other.

Prior studies have presented group polarization in various settings, ranging from political decisions (Janis 1989) to investment decisions (Whyte 1993). The role of group polarization, however, is mixed. Group polarization could be detrimental if group members lack profound understanding of it (Brockner 1992; McCauley 1989; Whyte 1993), yet it has been proved to be beneficial as well in many circumstances, such as shown in participation in social support systems (Festinger et al. 1956). Therefore, understanding what factors affect group polarization and its underlying formation process has become an important issue, helping people and decision makers to benefit most from group polarization.

As the progress of information technologies, the use of computer-mediated communication has shown to be able to make a difference in group polarization (El-Shinnawy and Vinze 1998). With computer-mediated communication, people engage in a group discussion with less social presence, compared with traditional face-to-face communications (Rice 1993). Seigel et al. (1986) present that dispersed group communications in an electronic form, which is with reduced social presence, have greater group polarization than face-to-face group communications. One explanation is given by Valacich et al. (1994): when people work together via computer-mediated communication in dispersed settings, more unique and high-quality ideas are generated, leading to greater group polarization. Sia et al. (2002), furthermore, investigates the effects of communication cues, social presence, and anonymity on group
polarization in computer-mediated communication. In particular, their studies have shown that when group members are in a distributed or anonymous environment, group discussions can lead to even higher levels of group polarization compared to traditional face-to-face meetings. This is attributed to two reasons. First, greater numbers of novel arguments are generated in computer-mediated communications, and second, there is higher incidence of one-upmanship behaviors.

While fruitful research has been conducted to investigate group polarization in both traditional face-to-face or computer-mediated communications in various settings, less is relatively known how group polarization works in one burgeoning form of computer-mediated communication, that is, communication in virtual communities, including online forum/message board communications. The innovation of the Internet technologies has encouraged more and more online forum and message board communications nowadays, changing how information is presented to people and the ways in which people can respond to that information. Take stock message board for example. Stock message boards, or stock forums, provide an unprecedented opportunity for investors to invest, debate, and exchange stock information and personal opinions with reduced or even no previous social connections. Statistics show that the level of forum activity keeps growing: during the year 1999 to 2001, more than 35 millions messages about public American firms were posted and discussed on Yahoo! Finance (Antweiler and Frank 2002). Stock message boards, in addition, are proven to have great impacts on the market. For example, Wysocki (1999) reports that message postings will forecast next day trading volume and next day abnormal stock returns. Although there is a noticeable importance and growing of stock message boards, the challenges and difficulties of examining communications happening in online forums are essentially rooted in three aspects. First, message board postings are known to have significant noise, contradictory opinions, rumors, and manipulations. The interactions among forum participants and online messages hence result in even more complicated communications. Second, the activity and scope of online forum communication are free of time and space constraints, leading to an enormous volume of information that can overwhelm forum participants. This also leads to the difficulty of identifying the form and boundary of communications among people. Third, the dynamic and complicated natures of online forums make it more difficult to identify factors and to develop measures for group discussions and communications, compared with other types of traditional or computer-mediated communications in prior work, where experiments can be more controlled.

I therefore attempt to unravel the understanding of group polarization in virtual communities through the study of one particular type of online forum: stock message boards. In this study, group polarization is recorded by comparing decisions at the group level and in online stock forum settings. Based on social comparison theory (SCT, Sanders and Baron 1997) and persuasive arguments theory (PAT, Kaplan 1977), which are considered two significant determinants of driving group polarization, I propose two factors, the group size (number of participants in one thread) and thread length (number of messages in one thread) to apply SCT and PAT to online communities setting. On message boards, although people do not necessarily know each other’s true identity, they do recognize other users’ virtual IDs, and as time goes by they take it more seriously their trust in other users when reading and responding messages from other participants they “virtually” know. In this way, forum participants build up a social connection to familiar others. The group size will therefore affect group decisions in the way that it decides how many people an individual can compare with, thereby recognizing whom they know and who they don’t know, assessing and adjusting his/her decisions, and leading to different beliefs and in turn choices. This explains how SCT is examined on stock message boards. Thread length, on the other hand, was intentionally to account of information novelty on message boards, which in part determines the persuasiveness of online messages. The idea is rooted in the fact that the longer the thread length, the more likely a novel message can be found, increasing the persuasiveness of a discussion thread and thus impacting group decisions. This relates to the PAT arguments.

To conduct the analysis, I examine the effects that group size and thread length have on group polarization. I first propose to use the “discussion thread” on stock message boards to represent a single discussion session. On stock message boards, a thread is a sequence of messages posted by a group of members, responding to one another. Each thread usually pertains to a single topic. During the interactions among people in a thread, a group discussion is considered to occur. Following Das and Chen (2001) and Gu et al. (2007), I then apply text mining approach to each message in one discussion thread to extract online investors’ sentiments, revealing online investors’ investment tendency pertaining to a particular stock. In other words, these sentiments represent online investor’s choices (decisions or positions) through communications, and thus are adopted as the index leading to group polarization. The
findings suggest two interesting results. First, group size affects group polarization on stock message boards, as suggested in SCT. Second, the effect of thread length depends on the posting activity of the message board; thread length has a significant effect on message boards with lower posting volume (lower level of activity) but does not have significant impacts on message boards with higher posting volume (higher level of activity). This explains how PAT works in online communities.

This study contributes to academics in two ways. First, telling choices or decisions from online forum users has always been considered a difficult task, given the dynamics and noisy contents on message boards. I shed some light on how to investigate group discussions and communications in online settings. More specifically, by proposing group size and thread length pertaining to discussion threads as two key factors that might impact online investors’ sentiments, I apply SCT and PAT for group polarization to the setting of virtual communities. Second, I adopt text mining approach for user sentiment extraction to represent forum participants’ choices. The results contribute to practitioners as well. Group polarization on online message boards refers to the situation that message board participants tend hold extreme opinions pertaining to an issue after participating on the message boards. This, if on stock message boards, can impact online investors’ decisions and in turn their trading behaviors and moreover future investments. Therefore, by understanding the factors that will or will not affect group polarization, decision makers such as message board providers can have a better understanding of the process of group polarization and if necessary, have control over it to gain more advantages.

The rest of the paper is organized as follows. In the next section, I provide a literature review on group polarization and computer-mediated communication. Specifically I will discuss two primary theories for group polarization, the social comparison theory and the persuasive argument theory. In the section of Methodologies and Hypotheses, I present the hypotheses and how the analysis of group polarization is conducted. In Data section, I describe the stock message board postings collected for the study. The model and results are given in Hypotheses Tests and Results section, followed by the discussions of limitations and implications of this study. I conclude the paper with conclusions and future work in the last section.

**Literature Review**

In the study of group communications and decisions, a group usually needs to reach a common decision even when the opinions of group members are in fact heterogeneous. For instance, Ackoff (1967) argues that in an organization, production and sales conclude a common strategy despite different respective goals. When group decisions are different from the original individual opinions, or when a group decision is enhanced after a group discussion, group polarization occurs. It refers to the situation that individuals tend to take more extreme opinions (Isenberg 1986; Myers and Lamm 1976). This phenomenon was first discovered by Stoner (1961) by finding out that group decisions are riskier than private decisions of individuals within the group.

Group polarization mechanisms could be investigated based on two primary theories, social comparison theory (SCT) (Sanders and Baron 1997) and persuasive arguments theory (PAT) (Kaplan 1977). SCT refers to the idea that individuals tend to assess themselves by comparing with other people. Furthermore, it is suggested by SCT that individuals will learn about and adjust their opinions toward the direction valued by other people as they continue the comparison process. The concept of SCT applies to group polarization in a way that by comparing with other people continuously, individuals tend to present themselves toward the socially desirable value after group discussions (Brown 1965). More specifically, Isenberg (1986) argues that one-upmanship and pluralistic balance are the two key characteristics that drive group polarization through social comparisons. First, one-upmanship, by definition, is the practice of trying to one-up, or outdo, an opponent. In a group setting, one-upmanship refers to the situation that when two individuals hold mutual positions during a group discussion, one tends to outdo another by moving his/her thinking toward the more socially desirable value. Pluralistic balance, on the other hand, refers to the “compromise” that people attempt to achieve, taking into account the tradeoff, or balance, between self-preference and preferences of other people. Therefore, during a group discussion, if an individual finds out that his/her position is not in accordance with the group value, which is called collective decision, he/she tends to move toward the collective decision, balancing between his/her own preference and the group’s. One-upmanship and pluralistic balance are found to be two important mechanisms that form the group polarization (Fromkin 1970; Pruitt 1971). In online communities, social comparisons keep occurring when forum participants interact with each other by reading and posting messages, and by arguing and
compromising through the interactions.

The second major theory, the persuasive arguments theory (PAT), is considered an informational influence (Kaplan 1977). According to PAT, individual decisions are determined based on how he/she weights the pro and con arguments. During group discussions, when each group member presents his/her pro and con arguments, he/she shifts the weight of these arguments, giving each other new arguments. Group polarization is then formed in the manner that people tend to pro the arguments that group discussions favor, and con those arguments against by the group discussions. This leads to the idea of to what degree each argument is considered to be “persuasive”. Novelty and validity are two proposed determinants of the persuasiveness. First, the novelty of an argument depends on how an individual regard it as interesting, original, and new. During the group discussion, when an individual is exposed to a novel argument, he/she tends to change the decisions. The second factor, validity, makes an argument to be more heavily weighted than a not-so-valid argument. During the group discussion, people will judge the correctness of each argument and change decisions accordingly. Prior studies have shown that both novelty and validity account of group polarization (Burnstein 1982; Butler and Crino 1992; Vinokur and Burnstein 1978). In the setting of online communities, it is considered harder to measure the persuasiveness of arguments. This is due to the common noisy contents and the alias mechanism used in online communities.

The approach to measuring group polarization is proposed from two different perspectives, at individual level as a preference change (Hinsz and Davis 1984) and at group level as a choice shift (Zuber et al. 1992). Choice shift refers to the difference between the individual's average “pre-meeting” choice and the final collective choice. When the final group decision is more extreme compared with the average pre-meeting decisions, group polarization occurs. In measuring choice shift, group polarization is considered at a group level. Preference change, which is measured for group polarization at an individual level, takes into account the average difference between an individual's pre- and post-meeting choices. After group discussions, if an individual's post-meeting choice moves toward the group choice, then group polarization is claimed to have happened. In the setting of online forum communications, the preference change is considered more difficult to capture. This is attributed to two reasons. First, in a virtual community, online participants come and leave arbitrarily, which means in a group discussion a member could appear and join a discussion but leave in the next second, without waiting till after the discussion. Second, the use of aliases in online forums makes it harder to position each different individual, leading to the challenge of identifying individual decisions. Therefore in this study, group polarization is measured at the group level, the group’s choice shift, based on Whyte (1993) as well, which shows that the magnitude of choice shift is generally stronger than that of preference change.

Computer-mediated communication is shown to impact group polarization (El-Shinnawy and Vinze 1998) with reduced social presence (Rice 1993). Siegel et al. (1986) show that in computer-mediated communications, where groups might communicate in a more dispersed environment, group polarization is intensified compared with the traditional face-to-face communications. Two explanations are given in prior work. First, when a group of members work together via computer-mediated communications, it is more likely that high-quality and interesting thoughts could be produced (Valacich et al. 1994). Moreover, in computer-mediated communications people are able to communicate in an anonymous way, which leads to a stronger group polarization because the exchange of social cues is restricted (Connoly et al. 1990) and social presence is again reduced. Sia et al. (2002) further presents the precise impact of features offered by computer-mediated communications, including communication cues, social presence, and anonymity, on group polarization and the underlying process. While there has been a growing interest in studying computer-mediated communications and group polarization, little is investigated about how one particular type of computer-mediated communication, online forums or message boards, affects group polarization. In the following sections, I will present an approach to understanding group polarization in online forums, specifically on the Internet stock message boards, extending prior work in computer-mediated communications and group polarization to a more dynamic and large-scale online setting.

**Methodologies and Hypotheses**

I conduct the study on group polarization analysis in the online settings of stock message boards. In contrast to the experimental designs suggested in prior work, for instance Sia et al. (2002), defining the scope of group discussions and communications in online forums has several challenges resulted from the
dynamics and complications on the Internet. Therefore, in my study, I first identify a discussion thread on stock message boards as a session of group discussion. A discussion thread consists of a sequence of messages posted by forum participants, responding to each other pertaining to a single topic. Figure 1 is an example showing what a discussion thread looks like.

In a discussion thread, forum participants read and post messages, responding to the arguments made by other people. This is analogous to the traditional group discussion settings: during a session of group discussion presented by a discussion thread, the posted messages are the arguments that people get involved, and by repeatedly interacting with one another, individual and group decisions are made.

As presented earlier, social comparison theory (SCT) and persuasive arguments theory (PAT) are the two major mechanisms that drive group polarization. Based on SCT and PAT, I propose two key factors that might affect group polarization in the stock message board settings: group size, the number of online participants in one discussion thread, and thread length, the number of messages comprising one discussion thread.

First, SCT refers to that forum participants tend to assess themselves, make decisions, and adjust their opinions by comparing with other online users. To what degree an online user can compare with others depends on one factor, that is, on how many people he/she can compare with, and that leads to the proposed idea of group size. On message boards, online users post and read messages, and interact with others using a virtual ID, an alias. Although the true identities of online users are generally unknown, for users frequently and actively participate in message boards, they in fact learn gradually about other users by interacting with each other. Forum participants thus do have social links, though virtually, to one another. Group size determines how many other online participants an individual can interact with, or compare with, in one discussion thread, leading to different beliefs and in turn different choices and decisions. In addition, social impact theory (Latane 1981) states that social influence experienced by an individual partially depends on the number of people involved, which in this scenario, refers to the group size of the discussion thread. The social impact theory implies that, the more people get involved in communications, the stronger impacts SCT and PAT processes have on each individual. This thereby results in stronger group polarization.

The second factor, thread length, is generated based on PAT. According to PAT, a forum participant’s decision is determined based on how he/she weighs the pro and con messages. The pros and cons of a message, furthermore, depend on the persuasiveness of the message. In particular, as discussed before, the novelty of a form posting is a significant determinant of its persuasiveness. It refers to how an online user regards a message as interesting, original, or new. The idea of thread length thus comes from the fact that having a larger number of messages will increase the possibility of including more novel arguments. This leads to the proposed factor, the thread length. Thread length determines how many messages contained in a discussion thread, and thus is considered a factor that can affect group polarization from the perspective of message novelty.

I next propose how to measure group polarization. Group polarization is widely measured using choice shift and preference change (Zuber et al. 1992). However, as discussed earlier, in online forum settings, the number and identity of forum participants are dynamically changing from time to time. Furthermore, prior study has also shown that the magnitude of choice shift is generally stronger than that of preference change (Whyte 1993). In my study, therefore, I focus on choice shift for group polarization. The choice of online investors is reflected by their “choices of stock investment,” which are expressed in the stock “sentiments” that online investors reveal in their posted messages. Notice that on stock message boards, for instance the case of Yahoo! Finance, online investors are usually given an option to specify

Figure 1. An example of a discussion thread (© 2007 Yahoo! Finance stock message board: MSFT).
their sentiments pertaining to a particular stock. This sentiment falls into five categories, including strong sell, strong buy, sell, buy, and hold. However, giving the sentiment is not mandatory, and very often forum participants do not specify it clearly. Therefore, in order to extract sentiments from messages posted by online participants, I follow prior work (Das and Chen 2001; Gu et al. 2007) and exploit text mining approach to mine the hidden sentiment information based on the textual message contents. In the text mining task, a training data set consisting of textual messages, each of which with a known, manually assigned sentiment class label, is first given. I then can induce a classification model from the training data to classify future messages into the pre-defined classes. For each discussion thread, posted messages are classified into one of the three categories in terms of their sentiments, shown as follows.

- **Category 0**: sentiment is “strong sell” or “sell”.
- **Category 1**: sentiment is “strong buy” or “buy”.
- **Category 2**: sentiment is “hold” or “no opinion”.

Given the three classification categories of sentiments, to measure the overall group choice and decision of a discussion thread, I use the measurement of sentiment entropy to capture the degree of how the sentiments are mixed, or the sentiment purity over all messages in one discussion thread. The sentiment entropy is calculated as follows. Assume in thread \( i \), the proportion of category 0 is \( P_0 \), the proportion of category 1 is \( P_1 \), and the proportion of category 2 is \( P_2 \). The sentiment entropy for each thread \( i \), is then calculated based on Shannon Entropy (Shannon 1948):

\[
\text{Sentiment Entropy } (i) = - P_0 \times \log_2 P_0 - P_1 \times \log_2 P_1 - P_2 \times \log_2 P_2,
\]

which is measure of how uncertain, or how mixed, the online participants’ decisions are. Therefore, in order to measure the degree of group polarization, I use the reciprocal of sentiment entropy. In other words, for each thread \( i \), the degree of group polarization is measured by

\[
\text{Group Polarization } (i) = \frac{1}{\text{Sentiment Entropy } (i)}.
\]

This is consistent with the idea that, the more extreme the individual decisions are, the smaller the sentiment entropy, and the greater the group polarization.

Based on the discussions above, I state the hypotheses as follows:

**Hypothesis H1. One stock message boards, group size will significantly affect group polarization.**

**Hypothesis H2. One stock message boards, thread length will significantly affect group polarization.**

Next I will present the data for the study, followed by the hypotheses tests and results to study the effects of group size and thread length on group polarization on stock message boards.

**Data**

In this study, I collected message board postings from a random sample of three stocks from Yahoo! Finance, which is widely acknowledged as one of the leading stock investment communities. The three stocks, presented by their stock tickers, include GE, MSFT, and MCD. The three stocks contain the issues that forum participants (individuals) care about and thus will have discussions over. According to the stock’s risk (i.e. beta) levels, these three stocks considered to be widely held and therefore less speculative. This characteristic is desirable because in the analysis of group polarization, we will want to exclude the influences of price movements and stock markets on the posted messages or on the online investors’ decisions. Based on the same reasons, for each of the three stocks, I collected data from the two weekend periods, which again aim to get rid of the stock price effects. For each stock message board during one weekend period, I collected messages from discussion threads which thread lengths are greater than or equal to 5. The following attributes were acquired for each message: thread topic, author...
ID, posting date and time, message content, and sentiment (if existent) assigned along with every message. Table 1 shows the summary of descriptive statistics of the discussion threads of all 3 stocks, including weekend periods, total number of threads, total number of forum participants over all threads, and the total posting volume as an index of the degree of posting activity and the frequency of participation of online investors on that particular stock message board. Table 2 further shows the descriptive statistics summary pertaining to the analysis of group polarization.

Table 1
Descriptive Statistics of Discussion Threads

<table>
<thead>
<tr>
<th>Stock Ticker</th>
<th>Weekend Periods</th>
<th>Total # of Threads</th>
<th>Total # of Participants</th>
<th>Total # of Postings (as of 05/01/2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE</td>
<td>04/07/07 ~ 04/08/07</td>
<td>38</td>
<td>108</td>
<td>210,063</td>
</tr>
<tr>
<td></td>
<td>04/14/07 ~ 04/15/07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSFT</td>
<td>04/07/07 ~ 04/08/07</td>
<td>54</td>
<td>186</td>
<td>341,689</td>
</tr>
<tr>
<td></td>
<td>04/14/07 ~ 04/15/07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCD</td>
<td>03/31/07 ~ 04/01/07</td>
<td>8</td>
<td>21</td>
<td>21,216</td>
</tr>
<tr>
<td></td>
<td>04/14/07 ~ 04/15/07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>04/21/07 ~ 04/22/07</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2
Descriptive Statistics for Group Polarization

<table>
<thead>
<tr>
<th>Stock Message Board Ticker</th>
<th>Thread Length Mean (Std. dev.)</th>
<th>Group Size Mean (Std. dev.)</th>
<th>Group Polarization Mean (Std. dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE</td>
<td>16.31 (9.95)</td>
<td>5.54 (1.94)</td>
<td>0.6716 (0.0309)</td>
</tr>
<tr>
<td>MSFT</td>
<td>15.75 (10.30)</td>
<td>7 (6.27)</td>
<td>0.6549 (0.0120)</td>
</tr>
<tr>
<td>MCD</td>
<td>7.75 (2.38)</td>
<td>3.28 (2.45)</td>
<td>0.6580 (0.0155)</td>
</tr>
</tbody>
</table>

Hypotheses Tests and Results

To perform statistical tests of the effects group size and thread length on group polarization, I first divide each of the factors into two levels: group size large vs. group size small, and thread length long vs. thread length short. The division threshold is based on the data we have. Figure 2 shows graphically the descriptive statistics of the discussion threads on GE and MSFT message boards (due to space constraints) after dividing factors into levels.
In this study, statistical tests are carried out based on 5% and 1% levels of significance. A 2 x 2 factorial ANOVA test involving the independent variable (group polarization) and two explanatory variables (group size and thread length, each containing 2 levels) is carried out. Table 3 shows the statistical results. F-test detects significant effect for group size on group polarization (GE: F = 38.72, p < 0.01; MSFT: F = 38.97, p < 0.01; MCD: F = 11.65, p < 0.05). Hypothesis H1 is supported. This result shows the consistency with SCT and the argument that group size significantly affects group polarization: group size determines how many other online investors an individual will be able to compare and interact with to adjust opinions and make decisions, thereby resulting in the phenomenon of group polarization.

For the effect of thread length, the results are mixed. For GE and MSFT, the F-test does not detect significant effect for thread length on group polarization (GE: F = 0.11, p < 0.05; MSFT: F = 0.90, p < 0.05). Hypothesis H2 is thus not supported: thread length does not have a significant effect on group polarization. This can be explained by the nature of online form messages. Recall that thread length was originally taken into account from the perspective that thread length determines the number of message included, relating to the possibility that a novel, persuasive message is found to occur. However, in online forums such as stock message boards, many of the messages are in fact small talks, containing noisy or irrelevant information. This impacts the effect of thread length: whether the discussion is long or short is no longer necessarily related to the persuasiveness of messages within it. This phenomenon is verified in my statistical test. For MCD, however, the F-test does detect significant effect for thread length on group polarization (MCD: F = 24.40, p < 0.01). This is attributed to the fact that MCD, compared with GE and MSFT, has relatively lower posting volume and thus lower level of posting activity (see Table 1. for total number of postings). The data also show that MCD has relatively shorter average thread length compared with GE and MSFT. This leads to the observation that, although many of the forum messages contain noise and are considered not persuasive, as discussed earlier, if there is only a smaller number of messages posted online, the importance of each additional message will still become more crucial, given that the original message pool is small. In other words, stock message boards with less posting volume and generally shorter threads will weigh the factor of thread length higher for group polarization than those with more posting volume and longer discussion threads.

Table 3
Statistical Results of 2 x 2 Factorial ANOVA

<table>
<thead>
<tr>
<th>Stock Message Board</th>
<th>Source of Variation</th>
<th>F</th>
<th>Sig. of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE</td>
<td>Group Size</td>
<td>38.72</td>
<td>0.000**</td>
</tr>
<tr>
<td></td>
<td>Thread Length</td>
<td>0.11</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td><strong>Adjusted R-Squared</strong></td>
<td>0.508</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Group Size</td>
<td>38.97</td>
<td>0.000**</td>
</tr>
</tbody>
</table>
Discussions and Limitations

In this study I examine the effects of group size and thread length on group polarization, in the setting of stock message boards. The results support the argument that group size significantly affects group polarization and the effects of thread length on group polarization depends on the posting activity of a particular stock message board. This study, however, has several limitations. First, this analysis does not account of the effects that the overall stock market performance has on individual and group decisions. For instance, although the discussion threads posted during weekend periods are taken into account, aiming to exclude the price effects, some external factors such as financial news could still impact online investors’ opinions pertaining to certain stocks (Tetlock 2007). The same argument exists for the limitation that in this study, only message boards of more stable stock are investigated. For volatile stocks, even though the analysis is performed during weekend periods, measuring group polarization still requires more careful considerations regarding additional factors resulted from the stock volatility that will impact group decisions. In addition, to enforce experimental controls and to reduce the complications on the stock message boards, this study was conducted using a relatively smaller number of stock message boards across 2 to 3 weekend periods. The numbers of forum participants and threads are sufficient for a reasonable analysis, but the phenomenon across a larger set of various stocks should be investigated more. Consequently, this study and findings open up an opportunity toward understanding group polarization in virtual communities, but will require more careful manipulations in terms of generalizability.

Several related aspects of academic implications can be undertaken. The analysis of group size and thread length effects on group polarization in this study can serve as a starting point for this line of research. Future studies can investigate additional factors that might as well impact group polarization. In particular, changes in the level of social presence, the degree to which people build personal connections with each other in a communication setting (Short et al. 1976), are shown to be able to affect group communications (Walton and McKersie 1965). Moreover, the exchange of communication cues, including verbal, visual, or textual cues (McGrath 1984), can change the level of social presence and human behavior (Johansen et al. 1991; Shorter et al. 1976; Sia et al. 2002; Williams 1977). Social presence can be observed in online forum communications based on the features such as participation time and frequency of online investors, and on how they interact with each other to establish the virtually social bonds. The way to capture the communication cues, however, is not easy. For online communication settings, it is natural to assume that only textual cues are present. The text mining approach provides a thought to manipulate the textual communication cues by filtering out noisy or irrelevant information and extracting the hidden yet informative knowledge from posted messages. This suggest an interesting future direction of research to further explore the effects of social presence and communication cues on group polarization, analogous with the study of Sia et al. (2002), in virtual communities.

Second, the concepts proposed in this study can be replicated in other online scenarios, for example, group communications and polarization in globally decentralized organizations. As more and more organizations are decentralized and organizational structures are getting flattened, many group decision makings in organizations nowadays are accomplished by means of other computer-mediated communication forms, for example, via not only emails but also weblogs or intra wikis (Wagner and Majchrzak 2007). Future studies can examine the underlying process and effects of group polarization in different virtual and cultural settings.

Results of this study also provide implications for practice. Decision makers for virtual communities can benefit from better understanding the process of group polarization and the effects that impact group polarization. Literature has shown that in some situations people benefit from a larger degree of group polarization, while in some cases more group polarization can do harms (Coleman 1957; Festinger et al. 1956). For virtual communities such as stock message boards, this understanding is particularly crucial.
because individual and group decisions can also lead to future investment, thereby impacting on trading volume, stock returns, revenues, etc. Online forum providers can therefore benefit from group polarization rather than being harmed by it by controlling the group polarization process.

Conclusion and Future Work

In this paper I examine the effects of 1) group size of a discussion thread and 2) discussion thread length on group polarization in virtual communities, particularly in the settings of stock message boards. The results suggest that group size has a significant impact on group polarization, while the effect of thread length depends on the degree of posting activity: thread length has a significant effect on group polarization for stock boards with lower activity level, but does not introduce a significant effect on group polarization for stock boards with higher posting activity. The findings echo SCT and PAT, which were proposed in prior literature for group polarization. In particular, SCT and PAT are applied to online communities and their effects are presented.

The increasing complexity of different forms in computer-mediated communications, such as virtual communities including online forums, message boards, weblogs, etc., has led to a growing interest in investigating the issue of group communications, decision making, and polarization, which can further impact critical organizational decisions and profits. The online communications, however, are known to be difficult to tackle with because of the dynamics and variability of the large-scale Internet platforms. This study is therefore considered as an incremental yet significant contribution towards a better understanding of how to measure group polarization in online communities, and of how group polarization has been affected.

This study also opens up several opportunities for academics and practitioners for future research directions. Some straightforward future work could be induced from the study. First, including more stock message boards and discussion threads from either stable or volatile stocks will help understanding the phenomenon of group polarization in online forums. Second, as discussed earlier, other factors such as social presence and communication cues can be examined regarding their effects on group polarization. In particular, one of the current progresses is to apply text mining approach to directly measuring the persuasiveness of online posted messages, comprising the discussion threads, and thereby affecting group polarization according to PAT.

References


**Micro-serendipity:**
Meaningful Coincidences in Everyday Life Shared on Twitter

Toine Bogers  
Royal School of Library and Information Science, Copenhagen, Denmark  
tb@iva.dk

Lennart Björneborn  
Royal School of Library and Information Science, Copenhagen, Denmark  
lb@iva.dk

**Abstract**

In this paper we present work on *micro-serendipity*: investigating everyday contexts, conditions, and attributes of serendipity as shared on Twitter. In contrast to related work, we deliberately omit a preset definition of serendipity to allow for the inclusion of micro-occurrences of what people themselves consider as meaningful coincidences in everyday life. We find that different people have different thresholds for what they consider serendipitous, revealing a *serendipity continuum*. We propose a distinction between *background serendipity* (or ‘traditional’ serendipity) and *foreground serendipity* (or ‘synchronicity’, unexpectedly finding something meaningful related to foreground interests). Our study confirms the presence of three key serendipity elements of *unexpectedness*, *insight* and *value* (Makri & Blandford, 2012), and suggests a fourth element, *preoccupation* (foreground problem/interest), which covers synchronicity. Finally, we find that a combination of features based on word usage, POS categories, and hashtag usage show promise in automatically identifying tweets about serendipitous occurrences.

*Keywords:* serendipity, Twitter, information behavior, information sharing, everyday life

**Introduction**

Serendipity has traditionally been defined as the accidental yet beneficial discovery of something one was not looking for directly, and has played an important role in many important scientific discoveries, such as x-rays and penicillin (e.g., De Rond & Morley, 2010; Merton & Barber, 2004; Van Andel, 1994). Serendipity also plays an integral part in everyday information behavior when “chance encounters with information, objects, or people [...] lead to fortuitous outcomes” (Rubin et al., 2011). As a consequence, methods and technologies for stimulating and supporting serendipity have received much attention in the field of information science. Technologies such as search engines, micro-blogging, and recommender systems have all been suggested as possible tools for increasing the potential for serendipity (see, e.g., McCay-Peet & Toms, 2011; Piao & Whittle, 2011; Zhang et al., 2012).

However, neither the study of the phenomenon serendipity nor the use of the concept in information science are without their difficulties. One problem is that there does not appear to be a single agreed-upon definition of serendipity: different definitions focus on different aspects, such as whether a serendipitous finding can be related to the active (foreground) information seeking task, or whether it has to be related to the background task alone. Different definitions assign different weights to personal and environmental factors. People also experience serendipity differently, have different thresholds for calling something serendipitous, and may use it synonymously with synchronicity, diversity, or novelty. What is needed, is a better understanding of the different ways people experience and communicate serendipitous occurrences in everyday life.
To address these issues, Erdelez (2004) called for approaches based on data generated by participants themselves to provide for more naturalistic studies of everyday serendipity. This was echoed by Rubin et al. (2011) who argued that most previous studies have been based on respondents’ descriptions elicited in interviews with researchers. They addressed this themselves by analyzing how bloggers describe their everyday serendipitous experiences (op.cit.). We follow up on their approach in this paper by using non-elicited, self-motivated user data from Twitter, the world’s largest online micro-blogging platform, to analyze how users share serendipitous experiences in the context of everyday life. We refer to this type of everyday serendipity as micro-serendipity.

Whereas Rubin et al. (2011) excluded blog posts that did not contain “a rich description including a mention of an accidental find and a fortuitous outcome”, we did not select user experiences in the present study based on a preset and fixed definition of serendipity. Instead, we want to understand what users themselves consider as serendipitous experiences and how they actually describe these experiences. We therefore deliberately omit a preset definition of serendipity in order to allow for the inclusion of micro-occurrences of what people themselves consider as meaningful coincidences in their everyday life.

With this research setting in mind, we address the following three research questions, which are part of an ongoing research project dealing with investigating everyday contexts, conditions, and attributes of serendipity as communicated on Twitter:

- What types of serendipity do Twitter users experience and communicate using Twitter? (RQ1)
- How often do people share serendipitous experiences on Twitter, and are there large individual differences in its frequency? (RQ2)
- What terminology do people use to describe and share serendipitous experiences on Twitter? (RQ3)

In sum, we see our main contribution as a detailed analysis of Twitter as a source for research into micro-serendipity as outlined above.

The remainder of this paper is organized as follows. The next section discusses our methodology with regard to data collection and annotation. The subsequent three sections (‘Experiencing serendipity’; ‘Frequency of serendipity’; ‘Describing serendipity’) address our research questions in turn, including both quantitative and qualitative analyses for each question. We discuss our findings and present our conclusions in the last section, along with plans for future work. Related work is discussed where directly relevant.

Methodology

Twitter has become a popular data source for social science research (e.g., Thelwall et al., 2011) with over 340 million tweets per day providing an unprecedented window into everyday life experiences, thoughts, interests, conversations, and language use of hundreds of millions of people. Our goal is to examine whether Twitter can be a suitable source for investigating serendipity as a broad, everyday phenomenon, as opposed to focusing solely on the role of serendipity in scientific discovery or information seeking. To this end, we crawled a set of 30,000+ English-language tweets containing the word *serendipity* spanning a period of seven months (August 2011 through February 2012). Details of this data collection can be found in the first subsection below. Even in a relatively focused crawl such as this one, the presence of the word *serendipity* does not guarantee that the tweeter is describing a personal serendipitous experience. To better understand the different ways people use the concept, we performed a content analysis of a subset of our data set, which we detail in the second subsection.

Data collection

While there are many different ways of describing serendipitous occurrences and experiences in 140 characters or less, we focus on tweets containing the word *serendipity* in order to limit the need for manual filtering, as well as in order to understand how people, who actually know the term, describe their serendipitous experiences. We explored the use of the Twitter API to collect our data set, but this is

---

1 http://blog.twitter.com/2012/03/twitter-turns-six.html
restricted to searching an index of between 6-9 days of the most recent tweets. Related work suggests that serendipity is a relatively rare phenomenon (e.g., André et al., 2009), so to be able to estimate how often individuals experience serendipity, we needed to collect tweets over a longer time frame. We therefore used Topsy, a search engine for content posted on Twitter. Topsy’s index contains tweets from as early as 2008, as we used Topsy to collect all tweets containing the word serendipity posted between August 1, 2011 and March 1, 2012. For this seven-month period Topsy contained 30,359 English-language tweets. Because Topsy is a text-based search engine, this automatically includes all 1,716 tweets that were tagged with the hashtag #serendipity. We refer to this 30,000+ data set as TOPSY-ALL.

One problem with using Topsy for our tweet collection is that the Twitter API places a limit of 1% of the total amount of public tweets that can be accessed using the API. This also influences Topsy’s indexing process. However, it is unlikely that the sample of Twitter that is indexed by Topsy, is biased towards or away from tweets about serendipity. We do not believe that these problems affect the conclusions we can draw from our data set collected using Topsy.

Coding tweets

To better understand how people tweet about serendipity, we performed a content analysis of a part of our data set and coded the tweets into different categories. To determine a list of appropriate coding categories, we took an open coding approach, where our coding categories emerge from the data (Lazar et al., 2010). Both authors developed their own coding categories based on a small set of 201 tweets published on February 1, 2012 taken from our original data set of +30,000 tweets, TOPSY-ALL. After calibration of our results, we merged our categories into a single coding scheme with five different categories:

- **COMM** tweets have a commercial intent, such as promoting jewelry, dresses or companies with the name Serendipity.
- **LINK** tweets contain links to Web content describing something related to the phenomenon serendipity.
- **NAME** tweets mention an object or location named Serendipity, such as movies, bars, restaurants, blogs, software, or bands.
- **REFL** tweets contain a general reflection, quote, or opinion about serendipity (but no clear description of a personal experience of a serendipitous occurrence).
- **PERS** tweets clearly describe a personal insight or experience of a serendipitous occurrence on the part of the tweeter.

These five categories are not mutually exclusive; a tweet about a personal serendipitous experience could be supplemented by a link describing the experience in more detail. Manually annotating all 30,000+ tweets in TOPSY-ALL was impractical, so we extracted two smaller data sets from our original data that we coded for our content analysis phase. Both of these data sets contain only tweets that have been tagged with the hashtag #serendipity. This is a result of observations made during the development stage of our coding scheme: tweets with the hashtag #serendipity contain a greater number of PERS tweets than tweets simply containing the term serendipity, and PERS is the category we are most interested in.

The first and smallest of the two additional data sets, is TOPSY-150, which contains the 150 tweets published during the first 11 days of February 2012, tagged with #serendipity. In order to test our five-category coding scheme, both authors annotated all tweets in this TOPSY-150 data set. Since our five categories are not mutually exclusive, using annotation reliability measures such as Cohen’s Kappa or Fleiss’ Kappa is not appropriate (Lazar et al., 2010). We therefore calculated inter-annotator agreement.

---

2 See https://dev.twitter.com/docs for more information.
3 http://www.topsy.com

4 A Twitter hashtag is a short user-generated keyword prefixed with the hash symbol (#) as a means of collating, sharing and following topics of interest in groups of users, who do not need to be connected through follower networks but take part in the same ‘hashtag streams’ (Bruns & Burgess, 2011). The same hashtag may be used for very different topics and events (as in our data set).

5 An alternative to Topsy could be the TREC Microblog track’s Twitter corpus (https://sites.google.com/site/treccorpora/). However, we do not believe this to be more representative of Twitter than the 1% of tweets Topsy has access to.

6 Only original tweets were coded as PERS; retweets were not considered to be personal serendipitous experiences of the person who retweeted the tweet in question. This is not 100% reliable as it is possible to hide the retweeted nature of a tweet.
(ITA) instead. Over all five categories combined, ITA is equal to 0.598. The ITA scores for the five categories are: LINK = 0.550; NAME = 0.633; COMM = 0.250; REFL = 0.414; PERS = 0.651. Coding tweets is a subjective task that is often open to debate. For some categories the relatively low agreement is due to the lack of context associated with many tweets, making it difficult to unequivocally determine the tweeters’ intent. However, for our main category of interest, PERS, we can consider the ITA as reflecting moderate agreement. After coding individually, we resolved any remaining differences through discussion to arrive at perfect agreement. When in doubt, we visited Twitter to inspect users’ tweet histories, links, and relevant tweet conversations to disambiguate the tweet in question. In our TOPSY-150 data set we ended up with the following category distribution: 1.3% of all tweets are classified as COMM, 27.3% as LINK, 39.3% as NAME, 18.0% as REFL, and 28.7% as PERS.

The next step in our annotation process was to analyze a bigger subset of TOPSY-ALL using our third data set, TOPSY-WINTER. This subset covers the three months of December 2011, January and February 2012 and contains 1073 tweets tagged with #serendipity. This means that the abovementioned data set TOPSY-150, which we used to test our coding scheme, is a subset of TOPSY-WINTER. In order to focus on our main category of interest (PERS) and to be able to code a larger set of tweets, we conducted a binary coding: tweets could be coded as PERS or as belonging to (at least) one of the other four categories (MISC). We did our best not to let established definitions of serendipity cloud our judgment; if a user clearly considered something to be serendipitous, then we marked it as PERS. After coding individually, we again resolved any remaining differences through discussion to arrive at perfect agreement. However, we were careful here not to overestimate: if a tweet was too ambiguous and/or contained too few contextual clues due to the 140-character limit of tweets, we assigned it to the MISC category. This resulted in the following for TOPSY-WINTER: 160 tweets (or 14.9%) fall into the PERS category and 913 (85.1%) in the MISC category. It is this TOPSY-WINTER data set that the majority of the work in the remainder of this paper is based on.

Experiencing Serendipity

In this section we present results aimed at answering what types of serendipity users experience and then communicate using Twitter (RQ1), and we examine three aspects related to RQ1 in greater detail: (1) whether serendipitous experiences are leisure-related or work-related, (2) the different thresholds people have for calling something serendipity, and (3) the differences and similarities between the closely related concepts of serendipity and synchronicity.

Leisure vs. work

In order to understand the activities and contextual situations that accompany and influence people’s serendipitous experiences, we performed a qualitative analysis of the 160 PERS tweets in TOPSY-WINTER, with a special focus on the distinction between leisure- and work-related activities. To increase our understanding of the context in the case of ambiguous tweets, we inspected users’ tweet histories on Twitter to examine the tweets surrounding the ambiguous tweet in question.

A total of 141 PERS tweets (88.1%) were coded as leisure-related and 14 tweets (8.8%) as work-related. One tweet was coded as both: I started typing ‘An Engineer’s Guide to Silicon Valley Startups’ into Google and ended up watching this: http://t.co/YgNRA3XZ #serendipity (id-J361). The work-related part about a startup guide is followed by a leisure-related link to a YouTube video found through Google’s auto-fill suggestions: “An Engineer’s Guide to Cat Yodeling (with Cat Polka)”.

Many work-related tweets dealt with meeting people that might lead to new business opportunities, such as: Just ran into @bruce_croxon, co-founder of lavalife. #serendipity (id-J332). The last four tweets were too ambiguous to be classified into either category.

A closer inspection revealed a rich diversity in leisure-related activities connected to serendipitous experiences. As shown in Table 1, the identified activities cover all kinds of digital and physical spaces in everyday life containing affordances for encountering “information, objects, or people” (Rubin et al., 2010).

---

7 In TOPSY-150, we checked 40 of the 150 tweets (or 27%) in this detailed manner.
8 For example: Rest in Peace Sarah Marie... 2/20/09 #Serenity (id-F273). In subsequent examples, ‘D’ stands for December, ‘J’ for January, and ‘F’ for February in our internally assigned tweet IDs.
9 The data sets used in this paper (TOPSY-150 and TOPSY-WINTER) are available at http://itlab.dbit.dk/~toine/?page_id=594. They contain Twitter IDs and Twitter user names for each tweet, our internally assigned IDs, and the result of our tweet annotation.
2011), including media, transportation, and shopping. The following examples may illustrate some of the rich diversity:

- **Nature**: First sight of cherry blossoms this year! #serendipity http://t.co/UFyMKhOM (id-F82). Link to an image of cherry blossoms in February.
- **Search engine + music (YouTube)**: Google found me a wrong Jung. And boy*, pleasant surprises and all. Almost a fan. Check http://t.co/Z6SFByo2 #Serendipity #guitarGeniouses (id-D41). Link to a YouTube video with the South Korean guitarist Sungha Jung.
- **People (chatting)**: Checked the weather and found an old friend! I'd say that means "plenty of sunshine"! #Serendipity @CindyManzanero http://t.co/AjKFWgQO (id-F280). Link to a smartphone screenshot of 'Baguio City Weekend Weather Chat'.

### Table 1
A sub-categorization of leisure activities connected to serendipitous experiences (N=160).

<table>
<thead>
<tr>
<th>Category</th>
<th>Activities</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media (total)</td>
<td>-</td>
<td>77</td>
</tr>
<tr>
<td>(Media: Articles)</td>
<td>article reading, quote finding</td>
<td>(12)</td>
</tr>
<tr>
<td>(Media: Internet)</td>
<td>blogging, chatting, podcasting, Facebook, Twitter, Google alert, search engine, web surfing</td>
<td>(21)</td>
</tr>
<tr>
<td>(Media: Music)</td>
<td>live, radio, Spotify, YouTube</td>
<td>(21)</td>
</tr>
<tr>
<td>(Media: Movie)</td>
<td>movie, tv</td>
<td>(5)</td>
</tr>
<tr>
<td>People</td>
<td>face-to-face, email, phone, lecture</td>
<td>22</td>
</tr>
<tr>
<td>Transportation</td>
<td>car, commuting, travel, wayfinding, nature</td>
<td>30</td>
</tr>
<tr>
<td>Shopping</td>
<td>money-finding, shopping, gifts</td>
<td>20</td>
</tr>
<tr>
<td>Food + Household</td>
<td>cooking, eating (incl. restaurant), cleaning</td>
<td>15</td>
</tr>
<tr>
<td>Sports</td>
<td>watching, performing</td>
<td>7</td>
</tr>
<tr>
<td>Others</td>
<td>photographing, artwork, ngo activism, religion</td>
<td>5</td>
</tr>
</tbody>
</table>

The next subsections contain more examples of micro-serendipity tweets in our TOPSY-WINTER data set.

### Serendipity thresholds

The qualitative analysis of the 160 PERS tweets in TOPSY-WINTER shows that there are clear differences in what could be called serendipity thresholds: when does a user find something unusual, unexpected, or surprising enough to consider it as serendipity? For example, this person links to an image of a pink balloon with the text 'Happy birthday princess': Found this balloon on the side of the road. How fitting. #serendipity http://t.co/e2kxdlhS (id-F228). The tweet mentioned in the previous section about the year's first sight of cherry blossoms (id-F82) is another example of how commonplace environmental occurrences can evoke serendipitous experiences. Being one of the few tweeters that posted more than one #serendipity tweet during the entire seven-month period covered by TOPSY-ALL (cf. section further below, 'Frequency of serendipity'), the same user also posted later on the same month: First rainbow I've seen this year. #serendipity (id-F173). These and similar observations in our data set indicate that different people have different thresholds for what they consider serendipitous; plain novelty or pleasant diversion may be enough. This finding echoes what Makri & Blandford (2012) emphasize about serendipity being a subjective phenomenon.

---

10 These categories are non-exclusive, i.e., some tweets cover two or more activities at the same time.
Serendipity vs. synchronicity

When Horace Walpole coined the word ‘serendipity’ in 1754 as “making discoveries, by accident & sagacity, of things which they were not in quest of” he stressed that “no discovery of a thing, you are looking for, comes under this description” (Van Andel, 1994 , p. 633).

In TOPSY-WINTER, 18.8% of our tweeters use the term serendipity when they, in unexpected ways, discover something they were already looking for, or in other ways were preoccupied with, such as: Walked into a mall I don’t know that well and choose the entrance right by the store I’m here for #Serendipity (id-D212). These kinds of experiences related to a preoccupied mind can also be seen in the following typical example: I joke to colleagues about feeling very Flashdance today & Spotify sends me: What A Feeling - Irene Cara #serendipity http://t.co/PfrzG6UG (id-J222).

It is clear from both our analysis of micro-serendipity and from related work (Makri & Blandford, 2012) that the notion of serendipity has broadened to encompass situations like the above. As noted by Makri & Blandford (2012), this kind of experience is also called synchronicity, which Wikipedia defines11 as “the experience of two or more events that are apparently causally unrelated or unlikely to occur together by chance, yet are experienced as occurring together in a meaningful manner.” Many of the serendipitous experiences involving music in the data set fall in this ‘synchronicity’ category, like in the Flashdance example above.

In the framework of the current research project we define synchronicity as a match between a perceived accidental occurrence in a person’s environment and a foreground activity, problem, or interest preoccupying that person12. In concordance with traditional definitions (cf. Merton & Barber, 2004; Van Andel, 1994), traditional serendipity may correspondingly be defined as a match between a perceived accidental occurrence in a person’s environment and a background interest that can be triggered in that person. We elaborate on this difference further below in the sections ‘Describing serendipity’ and ‘Discussion & Conclusions’.

Frequency of serendipity

Traditional studies of serendipity (e.g., André et al., 2009) have suggested that it is a relatively rare and anomalous phenomenon. While many people may have experienced serendipity, the same person may experience serendipity on a relatively infrequent basis. In this section we use our data sets to provide a realistic estimate of the frequency of serendipitous experiences shared on Twitter (RQ2).

One way of looking at this is from the perspective of Twitter as a whole: can we come up with a cautious estimate of what proportion of tweets describes serendipitous occurrences? According to Twitter, they processed 340 million tweets per day in March 2012 vs. 200 million in June 201113. A linear extrapolation14 of these numbers to the three-month period covering TOPSY-WINTER would mean an average of 305 million tweets per day in that period. English makes up approximately 51.5% of all tweets15 (Hong et al., 2011), which gives us about 157 million English language tweets per day.

In TOPSY-WINTER 160 out of 1073 tweets fall in the PERS category (14.9%). This amounts to an average of 1.8 PERS tweets per day that have been tagged #serendipity, as indexed by Topsy. Even if Topsy has access to just 1% of all Twitter’s tweets, this means that tweets describing serendipitous occurrences make up a vanishingly small proportion of the grand total of 157 million tweets. However, this is likely to be an underestimate, as we may assume that a considerable number of people tweet about their serendipitous experiences without tagging them #serendipity.

A better way of examining the frequency of serendipity may be within a personal context: how common is it for an individual to tweet about a personal serendipitous experience? If we look within our TOPSY-WINTER data set, we find that 146 different users account for 160 PERS tweets about a personal serendipitous experience using the hashtag #serendipity. Of these users, three sent out a pair of two identical tweets on different dates each. One other user retweeted the same tweet on eight different

12 This variant kind of serendipity has also been called pseudo-serendipity (e.g., Van Andel, 1994).
13 http://blog.twitter.com/2011/06/200-million-tweets-per-day.html
14 This is likely to be a conservative estimate given Twitter’s super-linear growth in the past as measured by the number of tweets per day.
15 This is assuming all languages grow at an equal rate, which is an oversimplified assumption.
dates, accounting for a total of 160 tweets. This means that in our three-month TOPSY-WINTER data set each user has only tweeted about a single unique serendipitous experience. This does not allow for any general conclusions to be drawn about the personal frequency of serendipity shared on Twitter. This could suggest that serendipity is a rare phenomenon—or perhaps a rarely communicated phenomenon—and that a larger data set is likely required to determine its frequency more reliably.

Describing serendipity

In this section we examine how tweeters describe their personal serendipitous experiences using Twitter: is there such a thing as a vocabulary for serendipity (RQ3)? We examine word usage, parts-of-speech, and Twitter hashtags associated with PERS tweets in our TOPSY-WINTER data set.

Word usage

In the first subsection below we make a quantitative analysis using a log-likelihood statistic on the data set. This is supplemented in the second subsection with a qualitative analysis of the data set.

Quantitative analysis. If we can determine the vocabulary commonly used to describe personal serendipitous experiences, it could help us identify other serendipity-related tweets in addition to the ones containing the word serendipity or the hashtag #serendipity. The presence or absence of such 'serendipitous' words could, for instance, be used as features for constructing a classifier that flags possibly serendipity-related tweets. In our situation we wish to determine whether certain terms are more characteristic for PERS tweets than for MISC tweets. This is similar, albeit on a smaller scale, to determining whether there is difference in word distributions between two text corpora. A robust measure for determining the surprise of a word's usage between two corpora is log-likelihood as proposed by Dunning (1993). Table 2 shows the 30 most characteristic terms in the TOPSY-WINTER data set for both the PERS and MISC tweets, ordered by log-likelihood.

<table>
<thead>
<tr>
<th>Tweet category</th>
<th>Top 30 most representative terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERS</td>
<td>just, found, road, met, book, walked, spring, ran, noticed, chinese, charity, car, note, today, song, store, simultaneously, shop, radio, pleasant, oooo, omg, immediately, bumped, picture, named, heard, flowers, flight, heard</td>
</tr>
<tr>
<td>MISC</td>
<td>watching, serendipity, hot, chocolate, movie, frozen, love, excited, welcome, clever sense, others, checked, create, beautiful, movies, discovery, spur, kate, chance, network, john, christmas, watch, search, panel, fave, york, sundae, heart, museums</td>
</tr>
</tbody>
</table>

A few of the PERS terms can be expected to be representative of tweets describing serendipitous experience, such as just, found, noticed, bumped, simultaneously, immediately, and omg (i.e., “oh my god”). The TOPSY-WINTER data set contains a lot of tweets mentioning the movie Serendipity, the Serendipity3 restaurant in New York, and blog posts about how innovation can lead to discovery and serendipity. This is reflected in the MISC terms, such as movie(s), kate/john17, frozen, chocolate, and discovery. Paradoxically, the word serendipity is much more indicative for MISC tweets than for PERS tweets due to its use in product names, which emphasizes the importance of finding other ways of detecting how people signal serendipitous experiences. In general, however, we believe TOPSY-WINTER to be too small for us to be able to discover a true vocabulary of serendipity unsupervised.

---

16 We extended the same analysis to our entire TOPSY-ALL data set, where a handful of these 146 users tweeted about a serendipitous occurrence more than once. The majority, however, did so only once in the seven-month period.

17 Kate Beckinsale and John Cusack are the lead actors in the movie Serendipity.
Rubin et al. (2011) compiled a list of 43 queries meant to identify blog posts about serendipitous occurrences, suggesting there is a certain terminology associated with serendipity. We received the full set of serendipity queries\(^{18}\) from Rubin et al. and treat them as another set of tweets to see if there is an overlap between the query vocabulary and the PERS and MISC tweet sets from TOPSY-WINTER. We calculate Kullback-Leibler-divergence (or KL-divergence) (Manning & Schütze, 1999) between the three different word distributions. KL-divergence is a measure of the similarity between two distributions; in our Twitter scenario, the lower the KL-divergence is, the more alike two text collections are in their word usage.

The KL-divergence between the PERS and MISC tweets is equal to 13.33. However, the KL-divergence between Rubin’s queries and the PERS tweets is equal to 20.31, whereas the KL-divergence between Rubin’s queries and the MISC tweets is equal to 20.25. The PERS and MISC tweets are thus much more alike in word usage than Rubin’s manually queries are to either of the two sets of tweets. In addition, there is virtually no difference between the KL-divergences of Rubin’s queries to the two sets (20.31 vs. 20.25). Rubin’s queries can therefore not be used to reliably distinguish between the two tweet categories and thus identify serendipitous tweets. This suggests that Twitter vocabulary is indeed different from the vocabulary used on blogs for describing serendipitous occurrences.

**Qualitative analysis.** Given the lack of clear results using log-likelihood, we took a closer qualitative look at the actual words signaling serendipity in the 160 PERS tweets in TOPSY-WINTER. Manually marking up these terms revealed a pattern of four key elements as shown in Table 3: preoccupation, unexpectedness, insight and value. The latter three elements were taken from a study by Makri & Blandford (2012), whereas preoccupation is a fourth key element we identified in our analysis. It is also discussed by Makri & Blandford (op.cit.), but not recognized as a key element.

<table>
<thead>
<tr>
<th>Preoccupation</th>
<th>continues (F123), I have been listening to ... almost exclusively for a week (D68), I was just (J288), not the first time (F381), read prior (J65), the same [...] (F41)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unexpectedness</td>
<td>actually (J351), and lo (J90), and look (F131), and then today (D68), bumped into (D137, J368), came across (J10), came on (F185), catch (D13), digging into (D35), dropped into (F102), find (J282, F28/118), found (D40/41/265 + 8 more), got it (J375), happen to have (D297), happens to be (J287), happened to get (J286), if not ... we would’ve missed (J357), just (D46/137 + 17 more), lucky (F165), met (J351, F360), omg (F343), popped up (D46), ran into (J332, F110/240), stumbled upon (J203), surprises (D41), there it was (D21), unexpected (F301), walked into (J152/267), while looking for something else (J10, F358)</td>
</tr>
<tr>
<td>Insight</td>
<td>and I see (D185), discover (F178), discovered (J424), found out (J287), haven’t seen it either (F149), hunch (F47), look at (J286), noticed (J344, F47/144), realized simultaneously (J414), immediately thought (F97)</td>
</tr>
<tr>
<td>Value</td>
<td>amazing (J152), appropriate (F332), awesome (D156), been looking for since (F310), cool (D165, J417), ,excellent (F48), 'i [exclamation mark] (F82 + many more), favorite (F85), fitting (F228), free (D276), good (D21, J147), great (J368/420), joy (D13), just in time (F291), love the web (D140), love it (J110, F364), love this (F93), pleasant (D41), score! (D156), smile (J147), :) [smiley] (F76), so perfect (F102), tasty (D148), timing perfectly (J394), what a lovely [...] (D109), whoa! (F120)</td>
</tr>
</tbody>
</table>

Table 3 shows that close inspection of the TOPSY-WINTER PERS tweets confirms the presence of all three key serendipity elements of unexpectedness, insight and value identified by Makri & Blandford, although not always explicitly so at the same time.

Sometimes just the element of unexpectedness could be identified with the elements of insight and value only implicitly present. This is illustrated by the following example: **Cool! Just when I was**

---

\(^{18}\) Example queries include “found some * * by accident” and “found * * by accident” (Rubin et al., 2011).
wondering what to get the wife for Valentine’s Day! #Serendipity http://t.co/5QibLEW (id-J417, link to image of dish drying rack). The element of unexpectedness is denoted by the word ‘just’ and the exclamation mark. Insight is shown by the link pointing to a humorous idea of a Valentine’s gift. The value is signaled by the word ‘Cool’, both exclamation marks, and the simple fact that the tweet has been shared on Twitter.

In another example: Realized simultaneously that I have no wrapping paper & the Chinese place left menu on my door. #serendipity #SorryDad (id-J414), unexpectedness is signaled by stating that, surprisingly, there is ‘no wrapping paper’ and that a solution to this problem is immediately and equally surprisingly found. The insight lies explicitly in the wording ‘realized’, whereas the implicit value is that the gift ended up being wrapped.

Our close inspection of the tweets suggested a fourth key element of serendipity: the user’s possible preoccupation (i.e., foreground problem/interest). This is in line with the results presented in the previous section (‘Serendipity vs. synchronicity’) about synchronicity. The following example illustrates how all four serendipity elements come together in some of the tweets: Found this balloon on the side of the road. How fitting. #serendipity (id-F228). As earlier mentioned, this tweet linked to an image of a pink balloon with the text ‘Happy birthday princess’. The preoccupation here is that the tweeter most likely knows a girl having birthday within a short time range. Unexpectedness is shown by the terms ‘found on the side of the road’. Both insight and value are expressed with the expression ‘how fitting’.

We manually compared the terms in Table 3 with the final set of 16 queries that Rubin et al. (2011) used for retrieving blog posts about serendipitous experiences. Their queries were permutations of the phrases “looking/searching for [...] but found/discovered”, “stumbled across/found [...] by chance/accident [...] looking for”, “wasn’t looking [...] but/when [...] found”, “found [...] while looking” and “accidently found”.

As shown in the previous subsection, there are differences with the terms used in our data set. A manual comparison showed that, while there is some overlap (e.g., ‘found’ and ‘stumbled’), Rubin et al. do not cover several of the variations in word usage present in our data set, e.g. ‘bumped into’, ‘came across’, ‘happened to get’, and ‘stumbled upon’. Twitter data revealing users’ actual word usage could thus suggest terms for other serendipity studies.

Parts-of-speech

Another aspect of the vocabulary of serendipity is which parts-of-speech (POS)—the lexical category of a word—are most often used to describe serendipitous occurrences. Whereas word usage tends to be much more varied, there is only a limited set of POS tags that can be assigned to a word, which could lead to clearer patterns in the description of PERS tweets. For POS tagging we used the MBSP toolkit (Daelemans & van den Bosch, 2005). We extracted the POS tags from our TOPSY-WINTER data set, for both the PERS and the MISC tweets, and filtered out stop words, symbols, and cardinals. This resulted in the following top 9 most frequent POS categories, shown in Table 4.

We find three interesting deviations from the normal distribution of POS categories. Nouns are about 10% more likely to occur in MISC tweets than in PERS tweets, which match our earlier observations that many tweets marked with #serendipity contain mentions of movies, bars, restaurants, and companies called Serendipity. In contrast, past tense verbs are twice as prominent in PERS tweets compared to MISC tweets. A likely explanation for this is that tweeting about serendipitous occurrences involves describing past events, necessitating the use of past tense verbs. Tweeting about watching, consuming or experiencing ‘resources’ called Serendipity are more often described in present tense as the event takes place, as evident from the 44% higher occurrence of present tense verbs. While none of these POS categories individually are enough to identify PERS tweets, these combined distribution patterns do show promise as features in a classifier for tweets about serendipitous occurrences.
Table 4
Distribution of the most frequent, non-stopword POS categories for PERS and MISC tweets.

<table>
<thead>
<tr>
<th>POS category</th>
<th>Relative frequency in PERS</th>
<th>Relative frequency in MISC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>34.80%</td>
<td>38.14%</td>
</tr>
<tr>
<td>Adjective</td>
<td>7.90%</td>
<td>8.85%</td>
</tr>
<tr>
<td>Adverb</td>
<td>7.35%</td>
<td>4.70%</td>
</tr>
<tr>
<td>Determiner</td>
<td>6.88%</td>
<td>4.68%</td>
</tr>
<tr>
<td>Pronoun</td>
<td>5.85%</td>
<td>6.13%</td>
</tr>
<tr>
<td>Verb, past tense</td>
<td>5.61%</td>
<td>2.83%</td>
</tr>
<tr>
<td>Verb, present tense</td>
<td>2.21%</td>
<td>3.19%</td>
</tr>
<tr>
<td>Verb stem</td>
<td>2.17%</td>
<td>3.62%</td>
</tr>
</tbody>
</table>

Hashtagging serendipity

A third way people can signal serendipitous occurrences is by using hashtags on Twitter. Our analysis of our TOPSY-WINTER data set shows that #serendipity is used by many users to describe serendipitous occurrences, but can we identify other hashtags that serve a similar purpose? A first step would be to identify the hashtags most frequently co-occurring with #serendipity for the PERS tweets in TOPSY-WINTER. Due to the relatively small size of TOPSY-WINTER and the fact that hashtags in general occur less often than normal words in tweets, the results of such an unsupervised identification are disappointing. The most commonly co-occurring hashtags typically represent locations and events, such as #nyc, #superbowl, #saints, and #weezercruise.

We can try to address the problem of data set size by looking at TOPSY-ALL: using log-likelihood we have determined the most characteristic hashtags that co-occur with #serendipity. An unsupervised approach does not seem to identify the most promising hashtags here either: #serendipity is often used to signal new content to a particular crowd or to describe the names of products or places, so the top of the list is dominated by such occurrences due to ‘spam’. However, a manual inspection of the list does reveal potential other serendipity-signaling hashtags, such as #synchronicity, #serendipitous, #chance, #insight, #wtf, #randomness, #accident, #lucky, and #surprise. These could be used in future work to collect additional tweets describing serendipitous occurrences. A fruitful method for future harvesting such tweets could for instance be scoring tweets by the number of these ‘signal’ hashtags they are tagged with and then select the highest-scoring tweets.

Discussion & Conclusions

In this paper, we have presented our work on micro-serendipity: investigating everyday contexts, conditions, and attributes of serendipity as shared on Twitter. One aspect of our investigation focused on how people share their serendipitous experiences on Twitter (RQ1). We found that users have very different thresholds for considering something serendipitous. We observe tweeters using the term serendipity for everyday occurrences ranging from pleasant diversions and distractions to wholly unexpected and unusual events. We propose generalizing this to a serendipity continuum to cover the entire spectrum of different degrees of surprise, from unplanned everyday incidents to unanticipated eureka moments in science. This is in line with Makri & Blandford (2012), who argue against viewing serendipity as a purely discrete concept.

While we cannot say anything conclusively about how often people experience serendipity in real life, we find that sharing such experiences via Twitter is relatively rare (RQ2). One explanation could be that serendipity is a rare phenomenon in general. However, people seem to have different thresholds for considering events serendipitous, which could mean that for some serendipity might be so common an everyday phenomenon that they do not always reflect upon it or find it worthwhile to share with others. Future work dealing on RQ2 would require annotating a larger collection of tweets taken at different points in time, possibly combined with an analysis of some individual hashtag streams (Bruns & Burgess, 2011) to determine which of these two explanations is most likely.
We do not have a conclusive answer to the question whether there is such a thing as a characteristic vocabulary of serendipity on Twitter (RQ3). On the relatively small scale of Topsy-Winter, the inherent variation of natural language dominates any patterns that might occur in the data. An analysis of the POS categories in PERS and MISC tweets revealed some interesting differences, whereas the analysis of hashtag usage was again inconclusive on its own due to the size of Topsy-Winter. However, while none of these three aspects of describing serendipity are characteristic enough on their own to detect tweets about serendipitous occurrences, we believe that combining these different serendipity-signaling features could be a promising approach to automatically detecting tweets about serendipitous occurrences. Tweets surrounding a suspected serendipitous tweet could also serve as extra context features here. Future work could involve training such a classifier for identifying such occurrences automatically. Techniques such as active learning (Manning & Schütze, 1999), where a small, annotated data set can be expanded in a continuous feedback loop of extracting features and detecting new patterns, may be useful here.

Our qualitative analyses revealed a pattern behind the diversified and punch-line word usage of tweets, confirming the presence of unexpectedness, insight, and value as identified by Makri & Blandford (2012). All three elements were present with different degrees of explicitation. Our study suggests that the person’s possible preoccupation could be included as a fourth serendipity element, thereby covering the aspect of synchronicity also present in the study by Makri & Blandford (op.cit.), but not included in their key elements. Just like the researchers interviewed by Makri & Blandford, the tweeters in our data set expressed a broad view on types of serendipity. In order to cover this range of serendipity types, we propose referring to the ‘traditional’ serendipity concept as background serendipity, because it is characterized by unexpectedly finding something meaningful related to a background interest, thereby changing that person’s focus and direction. The other frequent type of serendipity, not only experienced in everyday life, but also in science (op.cit.), is what we correspondingly refer to as foreground serendipity (i.e., synchronicity), as it is characterized by unexpectedly finding something meaningful related to a foreground interest and preoccupation, thus confirming the person’s focus and direction.

Both types of serendipity deal with people experiencing meaningful coincidences; in other words, with people considering an occurrence as both incidental and meaningful. This very consideration by the person matches what Makri & Blandford (op.cit.) call insight, and Horace Walpole called sagacity in 1754 (Merton & Barber, 2004; Van Andel, 1994). An occurrence must thus be considered as both incidental and meaningful in order for a person to denote it as serendipitous. Therefore, the two most important elements constituting a serendipitous experience seem to be unexpectedness and value—i.e., the meaningful coincidence—as preoccupation is not always current and some degree of insight must always be present in order to consider an occurrence as both unexpected/incidental and valuable/meaningful. Our study suggests that the respective thresholds for considering something as incidental and meaningful are highly subjective, explaining the range of the aforementioned serendipity continuum.

Summing up, Table 5 gives an overview of the key elements in background serendipity and foreground serendipity as discussed above.

Table 5

<table>
<thead>
<tr>
<th></th>
<th>Background serendipity</th>
<th>Foreground serendipity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preoccupation</td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td>Unexpectedness</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Value&lt;sup&gt;19&lt;/sup&gt;</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

In future work we will look more closely at how tweeters describe matches between environmental factors and their foreground/background interests. By casting serendipity as a correspondence between environmental and personal factors, we will extend our conceptual framework to include affordance theory as suggested by Björneborn (2010). In this framework, serendipity can be

<sup>19</sup> The elements of ‘unexpectedness’ and ‘value’ include ‘insight’ (Makri & Blandford, 2012) as discussed in the text.
seen as an affordance, i.e., as a three-way relationship between an environment, a human being, and a potential activity (cf. Dourish, 2004, p.118).

The work by Rubin et al. (2011) on analyzing blog posts that describe serendipitous occurrences served as an inspiration to the work described in this paper. While we have identified several similarities between the two approaches, there are also some key differences. By fine-tuning their queries, Rubin et al. aim for a precision-oriented approach to discovering anecdotal evidence of serendipity, based on a narrow, preset definition of the concept. In contrast, we cast a wide net and aim to identify more of the possible variations in the use of serendipity on Twitter, resulting in a data set spanning both information-related serendipity as well as everyday occurrences. While our data is more ambiguous due to the shorter length of tweets as compared to blog posts, and contains fewer quality descriptions of serendipitous occurrences according to the traditional definition, it does showcase the diversity in word usage and use of the concept to a much higher degree. Twitter data revealing users’ actual word usage could thus suggest terms for other serendipity studies.

Even if tweeters are afforded much less space for describing their inner thoughts than, for instance, bloggers, tweets do allow for more unfiltered (Bruns & Burgess, 2011), spontaneous, and near-instantaneous inspection, approaching real-time ‘streams of consciousness’ (as opposed to the retrospective nature of earlier serendipity research). If one exercises caution when filtering Twitter data, combined with close inspection as shown in our data analysis, we believe Twitter to be a promising resource for research into how people experience everyday life including micro-serendipity.

References


Abstract

There is an increasing number of Web data which consist of text and structured data, such as the combination of Wikipedia pages and DBpedia data. To issue queries to such data, we must choose one of the followings: (1) submit keyword queries against textual data part, or (2) submit structured queries written in structured query languages like SPARQL, against structured data part. Keyword queries are easy for casual users to write, but they do not have expressive powers enough to fulfill the user's information needs. On the other hand, structured queries are more expressive than keyword queries, but are not easy for casual users to write. This paper proposes a hybrid query language that seamlessly integrates the two types of queries, allowing us to write queries in a “pay-as-you-go” fashion.

Keywords: query languages, web search, structured data

Introduction

There is an increasing number of Web data which consist of text and structured data. An example is the combination of Wikipedia pages and DBpedia data (http://www.dbpedia.org/). DBpedia is RDF data that describe information appearing in the Wikipedia articles in a structured way. Another example is the combination of the Bible Ontology (http://home.bibleontology.com/) and text pages explaining the people who appear in the Bible1. Given the current activities such as Linking Open Data community projects (http://www.w3.org/wiki/SweolIG/TaskForces/CommunityProjects/LinkingOpenData) at W3C, we expect that in the near future it is common that many textual data have structured data (annotation) associated to them.

To issue queries to such data, you need to choose one of these types of queries: keyword queries against text pages, or structured queries, like SPARQL, against structured data. Keyword queries including Boolean queries are so simple and widely used by casual users. However, they cannot express complex conditions. In contrast, structured queries can express complex conditions, but are difficult for many users to write since it is required for users to know the complex syntax and schema-level information for writing queries.

This paper proposes Gradation query language (shortly, Gradation), a hybrid query language for the Web data that consist of text pages (Web pages whose main components are text) and structured data (mainly RDF data in the paper). In short, Gradation supports keyword queries with options to specify conditions on the associated structured data, and allows a seamless integration between keyword and structured queries. The design was developed based on an interesting fact that casual users often use

---

1 It is easy to associate the people that appear in texts with objects in the Bible Ontology.
“search options” in Web search, such as “site:edu.” In fact, a survey (internet.com K.K. (Japan), 2009) reported that about 25% of users of Web search engines have experience of using options.

The design of Gradation allows users to write not only both pure keyword and structured queries, but also hybrid queries that intermix conditions on text data and associated structured data. Thus, Gradation allows us to write queries in a “Pay-as-you-go” fashion, i.e., describe various types of queries according to the tradeoff between the acceptable cost of writing queries and the required preciseness of the query description.

Here is an example. When you search for text pages containing “Tom,” the query is:

\[ \text{Tom} \]

If the associated structured data contain the age of people, a query to get text pages containing “Tom” that correspond to those people who are over 40 is\(^2\):

\[ \text{Tom age} \geq 40 \]

As shown above, you can not only just submit simple keyword queries, but also add more complex conditions on structured data if required.

The contributions of the paper are as follows: First, we propose a unified query language that covers both simple keyword queries and complex structured queries. Next, we prove that the query language is relationally complete, because the relational completeness is a well-known criterion for discussing the expressive power of structured query languages. Finally, we show the experimental results to prove that that language allows us to write many hybrid queries in addition to pure keyword and relational queries.

In general, users need to know the schema-level information (e.g. attribute and class names) to write structured queries. An interesting application of our hybrid language is to help the user construct complex structured queries: The user first submits simple keyword queries and the system shows hints to transform the query into more precise ones. The detailed discussions on the application are beyond the scope of this paper and will appear in forthcoming papers.

The remainder of this paper is organized as follows: First, we describe related work. Second, we explain the processing model for Gradation query language. Third, we explain Gradation queries. Next, we give the formal semantics and prove that Gradation is relationally complete. Then, we show some experimental results to illustrate Gradation allows us to write many queries other than pure keyword queries and relational queries. Finally, we describe the conclusion.

### Related Work

Since Gradation is a hybrid query language for text and associated structured data, there are a lot of related work and tools that address queries for text and structured data in different ways. They include: (1) natural language queries for structured data, (2) keyword queries for structured data, (3) structured query languages that have built-in keyword search mechanisms, (4) full-text search engines that allow users to specify conditions on structured metadata and (5) sophisticated search systems.

**1. Natural language queries.** It would be ideal if users can write queries in natural languages. Therefore, there have been a lot of attempts (Androutsopoulos, Ritchie, & Thanisch, 1995). A typical approach is to translate natural language queries into structured ones. However, it is difficult for computers to perfectly understand what users intended with the queries, and this is one of the reasons why we have artificial and formal query languages including SQL, SPARQL, and Gradation. Essentially, processing natural language queries is to map such queries to well-defined formal abstractions. In that sense, researches on natural language query processing and formal query languages are complementary to each other.

**2. Keyword queries for structured data.** There are many researches on keyword queries for structured data (Chen, Wang, Liu, & Lin, 2009). For example, SPARK (Luo, Wang, & Lin, 2008) is a keyword search engine against structured data, which allows users to mention attributes or table names in queries. SPARK is different from Gradation in that (1) it is a query language for structured data only, and (2) the relationships among tuples the queries can represent are limited thus SPARK is not relationally complete. Another approach is to try to find the semantics of keyword queries in the context of given structured data (Sarkas, Paparizos, & Tsaparas, 2010), in which they try to find which attribute corresponds to each given keyword. Again, Gradation is different from such researches, in that it does not try to find the intention of keyword queries, and is complementary to them.

**3. Structured query languages with keyword search mechanisms.** XQuery and XPath Full-Text (Case et al., 2011) extends the syntax and semantics of XQuery and XPath to realize full-text search on XML. Many SQL databases support full-text search functions. Semplore is an IR-based system that has a query language for text and structured data. Because they are extensions of structured languages or a novel query language, users cannot write even simple queries without knowing the syntax. In contrast, Gradation is a natural extension of keyword queries, and allows us to write queries in a “Pay-as-you-go” fashion.

\(^2\)You can quote character strings (attribute values, or keywords) to distinguish them from reserved words or attribute names if necessary.
(4) Full-text search engines allowing queries on structured metadata. An example is Lucene (http://lucene.apache.org/core/), a search engine given as a Java library to develop full-text search systems. Programmers, who use Lucene to implement a search engine, define queryable fields for given text data in advance, and then users of the search engine can use advanced search functions to specify the fields. Unlike Gradation, Lucene does not define how to write queries since it is not a query language. In addition, the functions for structured search given by Lucene are limited and do not provide the relational completeness.

(5) Sophisticated search systems. There is an increasing number of sophisticated search systems, such as faceted search systems (Tunkelang, 2009) and fielded search systems (e.g., LexisNexis (http://www.lexisnexis.com/en-us/)). Gradation is different from the systems in that it is a relationally complete language that seamlessly integrates keyword search and structured queries.

In addition, the ranking of query results is an important issue of in Text retrieval and Web search (Selvan, Sekar, & Dharshini, 2012). Gradation has a set-based semantics and is neutral on the ranking issue. The development of ranking schemes for Gradation is an interesting future work.

Query Processing Model

This section explains the query processing model for Gradation. We first define the data model of Gradation, and then, explain the input and output of the query processing model.

The Data Model

We define the data $D$ as a triple: $D = (T, G, f)$. Figure 1 illustrates an example of $D$.

- $T$ is a set of text pages. For example, $T = \{\text{Samurai}_t, \text{M:I}_t, \text{ToyStory}_t, \text{Ken}_t, \text{Tom}_t, \text{Johnny}_t\}$ is in Figure 1.

- $G$ is the RDF data associated to $T$. RDF (Resource Description Framework) (http://www.w3.org/RDF/) is a framework for describing metadata. RDF describes metadata in terms of a set of resources, a set of values, and the relationship among them. RDF data consists of a set of triples, each of which has the form of $(\text{Subject}, \text{Predicate}, \text{Object})$, in which Subject is a resource, and Object is either a resource or a value. For example, given a resource Tom_r, assertions on Tom_r such as “Tom_r is 50 year-old” or “Tom_r belongs to Actor class” can be described using a set of triples: $\{(\text{Tom}_r, \text{age}, 50), (\text{Tom}_r, \text{type}, \text{Actor})\}$. Given a set of RDF triples, we can express it as a labeled directed graph $G = (N, P, E)$ (Figure 1 (top)). Here, $N$ is a set of nodes, $P$ is a set of predicates, and $E$ is a set of edges. Each edge in $E$ is a triple $(n_i, p, n_j)$ where $n_i, n_j \in N$ and $p \in P$. Triples of RDF data are mapped to edges, in which resources and values are mapped to nodes, and predicates are mapped to edge labels. In Figure 1, $N = \{\text{Tom}_r, 48, \text{ToyStory}_r, \text{Film}\}$ and $E = \{\text{(Tom}_r, \text{age}, 48), (\text{ToyStory}_r, \text{type}, \text{Film})\}$.

- $f : T \rightarrow N$ is an injective function that expresses the relationship between text pages in $T$ and nodes of $G$. If $t \in T$ is associated to $n \in N$, the fact is represented by $f(t) = n$. For example, $f(\text{Tom}_t) = \text{Tom}_r$ in Figure 1.

Inputs and Outputs

Figure 2 illustrates the query processing model for Gradation. The model consists of the following components: (1) $D$ is the data explained in the previous section, (2) $q$ is a submitted query, (3) $R$ is the relation to represent the result of applying $q$ to $D$.

The schema of relation $R$ is determined by $q$. In the simplest case, $R$ is a unary relation whose only attribute is “t_uri” which means the URI of the text pages. For example, if $q$ is “Tom”, $R$ is a unary relation contains a set of URIs of text pages containing “Tom”. Other cases will be explained in the next section.

Gradation Query Language

This section explains Gradation queries with examples. We start with simple queries that appear in casual searching and gradually show more precise ones. Therefore, this section illustrates the flexibility of the language. All of the examples take as input the data shown in Figure 1. The purpose of the section is to give intuitive explanation. The formal semantics are given in the next section, and the detailed syntax will be shown in the Appendix.

---

3 Here, strings ending with "_t" denote URIs of text pages.
4 Here, strings ending with "_r" denote URIs of resource nodes.
5 "type" will be used to refer to "rdf:type".
A query consists of keywords and non-keyword components (Table 1). Non-keyword components are used to write more complex queries that require path expressions, class specifications, and arithmetic comparisons. Gradation allows us to write queries with various combinations of keywords and non-keywords components, covering both of simple keyword queries and complex structured ones. The user can choose appropriate combinations according to the tradeoffs between the acceptable cost of writing queries and the required preciseness of query description. In the followings, we first explain queries consisting only of keywords. Then, we explain queries having non-keyword components.

Queries with One Keyword

The section explains queries consisting of only one keyword, and their semantics. In Gradation, a query to get the text pages containing “Tom” can be written as:

Tom (Q1)

As another example, a query to get the text pages containing “Samurai” is:

Samurai (Q2)

The results of the queries Q1 and Q2 are shown in Figures 3 and 4, respectively.

In general, if a query consists of one keyword \(k\), the query processor performs a Boolean full-text search against \(T\) with \(k\), and the result is a set of URIs of text pages containing \(k\). Formally, the result is \(\{t | t \in T, \text{match}(t, \text{keyword})\}\).

Gradation supports the wildcard ("*") as a special type of keyword. The result of the query consisting only of "*" is \(T\), namely, the set of all the text pages. The wildcard can be used at the place in which predicates or string values are required (i.e., at places in which class names, attribute names, or string values are required).

Intersection, Union and Difference

In order to make a query to produce the intersection, union, and difference of the results of subqueries (Lines 1–3 of Table 1), users use non-keyword components and, or, and -, respectively. For example, a query to get the text pages containing “Tom” and “Samurai” is:

Tom and Samurai (Q3)

The result of query Q3 is shown in Figure 5. This is the intersection of the results of Q1 and Q2 (see Figures 3 and 4). In Gradation, the intersection is the default interpretation to connect subqueries. Therefore, and can be omitted and the query “Tom Samurai” is interpreted as “Tom and Samurai”.

Figure 1. Example of \(D\)

Figure 2. Query Processing Model of Gradation
Table 1
List of Non-keyword Components of Queries

<table>
<thead>
<tr>
<th>Query with a non-keyword component</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_1 \text{ and } q_2 )</td>
<td>Intersection. Returns the intersection of the results of queries ( q_1 ) and ( q_2 ).</td>
</tr>
<tr>
<td>( q_1 \text{ or } q_2 )</td>
<td>Union. Returns the union of the results of queries ( q_1 ) and ( q_2 ).</td>
</tr>
<tr>
<td>( q_1 \text{ - } q_2 )</td>
<td>Difference. Returns the set of tuples that appear in the result of ( q_1 ) and do not appear in that of ( q_2 ). If ( q_1 ) is omitted, &quot;<em>&quot; is used for ( q_1 ) (Here, &quot;</em>&quot; is a wildcard for keywords. See Section “Queries with One Keyword” ).</td>
</tr>
<tr>
<td>( \bigcup ( q_1 \ldots q_i \ldots q_n ) )</td>
<td>Grouping. ( q_1 \ldots q_i ) are evaluated before the others.</td>
</tr>
<tr>
<td>( \text{class: } c )</td>
<td>Class Instances. Here, ( c ) is a class name in RDF data. The query returns the set of instances of class ( c ), namely the set of resource nodes have ( \text{rdf:type} ) predicate connected to ( c ) or to a class reachable from ( c ) through subclass properties.</td>
</tr>
<tr>
<td>( q_1 \text{ [ attribute1, \ldots, attribute_n ] } )</td>
<td>Make Relational Attributes Appear in the Result (Projection). Here, ( \text{attribute}_i ) is one of the followings: (1) ( \text{t_uri} ), which means the URI of text pages, (2) ( r_\text{uri} ), which means the URI of resource nodes, and (3) a predicate that appears in the given RDF data.</td>
</tr>
<tr>
<td>( q_1 \cdot p \cdot q_2 )</td>
<td>Path Traversal. Let ( N ) be a set of RDF graph nodes corresponding to the result of ( q_i ). The query returns the set of ((t_1,t_2)), in which ((n_1,p,n_2)) ( \in ) ( E ), ( n_i \in N ), ( n_i = f(t_i) ) and ( t_i \in T ).</td>
</tr>
<tr>
<td>( \text{attribute}_1 \theta \text{ value or } \text{attribute}_2 \theta )</td>
<td>Attribute-based Selection. This component returns the set of URIs of text pages each of which is connected to RDF graph node ( s ) satisfying one of the following conditions: (1) There exist ((s,\text{attribute}_1,\text{value})) ( \in ) ( E ) s.t. ( \text{attribute}_1 \theta \text{value} ) holds. (2) There exist ((s,\text{attribute}_1,o_1),(s,\text{attribute}_2,o_2)) ( \in ) ( E ), s.t. ( o_1 \theta o_2 ) holds. For ( \theta ), we can use ( &gt;, &lt;, &gt;=, &lt;=, =, ==, ) or ( != ).</td>
</tr>
<tr>
<td>( q_1 \ast q_2 )</td>
<td>Cartesian Product. The query returns the Cartesian Product of ( q_1 ) and ( q_2 ) (Note that each query returns a relation in the Gradation query processing model.)</td>
</tr>
<tr>
<td>( q_1 \text{ as Sa q}_2 \ldots q_n )</td>
<td>Alias. In ( q_2 \ldots q_n ), ( S_a ) can be used as the name to refer to the result of ( q_1 ).</td>
</tr>
</tbody>
</table>

Figure 3. Output of Q1 Figure 4. Output of Q2 Figure 5. Output of Q3

Grouping of Subqueries
Users can use \((...)\) to group subqueries (Line 4 of Table 1), to affect the evaluation order of subqueries. For example, the following queries have different evaluation orders of subqueries.

\[
\begin{align*}
\text{(Tom or Ken) Scientist} & \quad \text{(Q4)} \\
\text{Tom or (Ken Scientist)} & \quad \text{(Q5)}
\end{align*}
\]

Here, Q4 returns the intersection of the set of text pages containing “Tom” or “Ken,” and the set of text pages containing “Scientist.” On the other hand, Q5 returns the union of the set of text pages containing “Tom,” and the set of text pages containing “Ken” and “Scientist.”

Restricting to Instances of a Particular Class
Gradation has a non-keyword component \( \text{class: } c \) in order to restrict the query results to instances of a particular class \( c \) (Line 5 of Table 1). For example, a query to get the set of text pages corresponding to Tom who is an actor is:

\[
\text{Tom class: Actor} \quad \text{(Q6)}
\]
The result of Q6 is shown in Figure 6. The semantics of the query is the intersection of the results of subqueries Tom and class:Actor (Figure 3 and 7). In general, the result of class:c is the set of text pages corresponding to resource nodes which are instances of c. Here, we say a resource node is an instance of c if the node has rdf:type predicate connected to c or to a class reachable from c through subclass properties in the RDF data G. Therefore the subquery class:Actor returns the set of URIs of text pages that corresponds to RDF resource nodes having rdf:type predicate connected to Actor class or to a class reachable from Actor class through subclass properties in the RDF data G.

Specifying Relational Attributes to Appear in the Output

Each Gradation query can have a sequence of attribute names following the query (i.e., q[attribute_1, ..., attribute_n]) in order to provide each relational attribute with a tag “to appear” meaning that the attribute appears in the final result (Line 6 of Table 1). When the attributes correspond to RDF predicates, the attributes will contain values in the given RDF data G. For example, a query to return a set of binary tuples with (1) URIs of text pages that contains ‘Tom’ corresponding to resources who are of actor type, and (2) the values of their ages is:

(Tom class:Actor)[t_uri, age](Q7)

Here, t_uri is the attribute to contain URIs of test pages (explained next). The result of Q7 is shown in Figure 8. The query returns a relation which has two relational attributes, t_uri and age.

Here, attribute_i can be one of the followings:

1. t_uri: URIs of text pages that are contained in the result of q.
2. r_uri: URIs of resource nodes corresponding to the values of t_uri (i.e., r_uri = f(t_uri)).
3. Predicate p (e.g., age), that is connected to the RDF nodes in r_uri. If the RDF nodes do not have p, the values of the relational attribute of the final result will be null.

There are two points to note. First, t_uri's are tagged as “to appear” by default (i.e., query q without [attribute_1, ..., attribute_n] is interpreted as q[t_uri] by default). For example, query “Tom” is interpreted as “Tom [t_uri]”. Second, the expression can be nested in queries and the tags are overwritten by the outer ones (i.e., if an attribute that is given “to appear” tag by an inner one is not specified in the outer one, the tag is removed from the attribute). For example, the following nested queries (Q8-Q10) return the results in Figures 9-11.

((Tom class:Actor)[t_uri, age])[t_uri, age](Q8)
((Tom class:Actor)[t_uri, age])[t_uri](Q9)
((Tom class:Actor)[t_uri, age])[age](Q10)

Traversing Paths in the RDF Graph

Gradation allows users to traverse the paths in the given RDF data (Line 7 of Table 1). For example, a query to get the set of pairs (binary tuples) of text pages, in which each tuple has a text page corresponding to a film and a text page corresponding to an actor who stars in the film, is:

class:Film.starring.class:Actor(Q11)
The result of Q11 is shown in Figure 12.6.

In general, query \( q_1, p.q_2 \) returns a set of binary tuples, each of which represents a one-length path to traverse a link to connect the results of queries \( q_1 \) and \( q_2 \). In other words, let \( T_1 \) be the set of \( t_{\_uri} \) values of the result of \( q_1 \), and let \( N_1 = \{n|n = f(t), t \in T_1 \} \). Then, the result of \( q_1, p.q_2 \) is a set of binary tuples \( \{(t_1, t_2)|\forall i \in \{1, 2\}(n_i \in N_1, n_i = f(t_i)), (n_1, p, n_2) \in E\} \). Since the default attribute for the result is \( t_{\_uri} \), the query \( q_1, p.q_2 \) is interpreted as \( q_1, p.q_2[t_{\_uri}()\to t_{\_uri}()\to] \) if not explicitly specified.

For example, consider query Q11 and assume that \( N_1 = \{M:I_t, Samurai_t, ToyStory_r\} \) and \( N_2 = \{Tom_r, Ken_r, Johnny_r\} \) are the set of instances of Film class and that of Actor class, respectively. Then, the result of Q11 is \( \{(M:I_t, Tom_t), (Samurai_t, Tom_t), (Samurai_t, Ken_t)\} \).

As a natural extension, users can write arbitrary length of path traversal \( q_1.p_1.q_2.p_2.q_3.\ldots.p_{n-1}.q_n \). In this case, let \( N_i \) be the set of RDF nodes that are \( t_{\_uri} \) values for the result of \( q_i \). Then, the result of the query is a set of \( m \)-ary tuples \( \{(t_1, t_2, \ldots, t_m)|\forall 1 \leq i \leq m(n_i = f(t_i)), \forall 1 \leq i < m(n_i \in N_i, (n_i, p_i, n_{i+1}) \in E)\} \).

Combining Keyword Queries with Path Traversals

This section shows an example of a query having a keyword and a path traversal:

\[
\text{Tom class:Film.starring.class:Actor} \\
\text{(Q12)}
\]

Note that the query is an example of a hybrid query (neither a pure keyword nor a pure relational query). A subtle point here is that the results of \( \text{Tom} \) and \( \text{Film.starring.class:Actor} \) are not union compatible, because the former returns a unary relation and the latter returns a binary relation. Therefore, we cannot compute the intersection of the two results. However, as will be explained in a later section (Resolution Rules for Union Incompatibility), Gradation has a resolution rule for the case in which one subquery is a keyword query. Let \( R_1 \) is the result of a keyword subquery (with a keyword \( k \), such as \( \text{Tom} \)) and \( R_2 \) is the non-unary relation computed by the other query \( q_2 \). In this case, the result of \( k \) and \( q_2 \) contains a non-unary tuple \( t \in R_2 \) if a text page addressed by one of the \( t_{\_uri} \) attribute values of \( t \) contains \( k \).

For example, let \( R_{\text{path}} \) be the result of the second subquery of Q12, and assume that \( R_{\text{path}} = \{(M:I_t, Tom_t), (Samurai_t, Tom_t), (Samurai_t, Ken_t)\} \). Then, since every tuple in \( R_{\text{path}} \) has at least one URI of text pages containing “Tom,” the result of Q12 is the same as the result of Q11 (shown in Figure 12).

Applying Grouping in Various Types of Queries

Users can apply grouping (explained in an earlier section (Grouping of Subqueries)) to various types of queries including non-keyword queries. For example, a query to get the set of pairs of text pages, in which each binary tuple has (1) a text page containing “Tom” that corresponds to an actor and (2) a text page for a film in which he stars in is:

\[
\text{class:Film.starring.(Tom class:Actor)} \\
\text{(Q13)}
\]

The result of Q13 is shown in Figure 13. As with the case of keyword queries, the subqueries in () is evaluated before the other subqueries. In Q13, the subquery \( \text{Tom class:Actor} \) is evaluated first and returns the set of URIs of text pages having “Tom” and corresponding to the resources are actor type. Therefore, the result of Q13 (shown in Figure 13) is different from that of Q12 (shown in Figure 12), and does not have the tuple \( \text{Samurai_t, Ken_t} \).

Selection Based on Attribute-Values

The following is an example of a query to get the URIs of text pages corresponding to actors who are over 50:

\[
\text{class:Actor age>=50} \\
\text{(Q14)}
\]

---

6 If at the output relation has different attributes with the same attribute names, the names would be changed appropriately.
The result of Q14 is shown in Figure 14. Attribute-based selection (Line 8 of Table 1) can be specified as (1) \( \text{attribute}_1 \theta \text{value} \) or (2) \( \text{attribute}_1 \theta \text{attribute}_2 \) where value is a constant. Here, the former is the comparison between values of two attributes and the latter is the comparison between values of an attribute and a constant. The output of \( \text{attribute}_1 \theta \text{value} \) is the relation that contains URIs of text pages each of which is connected to RDF graph node \( s \) s.t. \( (s, \text{attribute}_1, \text{value}) \in E \) (See an earlier selection (The Data Model) ) and \( \text{attribute}_1 \theta \text{value} \) holds (for example, when \( q \) is "age>=50," \( s = \text{Ken}_r \)). Similarly, the output of \( \text{attribute}_1 \theta \text{attribute}_2 \) is the relation that contains URIs of text pages each of which is connected to RDF graph node \( s \) s.t. \( (s, \text{attribute}_1, o_1), (s, \text{attribute}_2, o_2) \in E \), and \( o_1 \theta o_2 \) holds (For example when \( q \) is "birthPlace==activePlace," \( s \) can be \( \text{Ken}_r \)).

Alias

Gradation allows users to give alias names to subqueries (that returns relations) so that they can write complex queries, such as queries having the same path traversal multiple times (Line 9 of Table 1). For example, assume we want to know the films such that (1) two actors star in the same film, (2) one of them has the text page that contains "Tom" and (3) the other has the text page that contains "Ken." Then, the query is:

\[
(\text{class:Film as $a$}).\text{starring.}(\text{class:Actor Tom}) \ a.\text{starring.}(\text{class:Actor Ken})
\]

The result of Q15 is shown in Figure 15. It has three attributes: (1) the URIs of text pages containing "Tom" and corresponding to RDF nodes of actor type, (2) the URIs of text pages containing "Ken" and corresponding to RDF nodes of actor type, and (3) the URIs of text pages corresponding to films they star in, because the \( t\_\text{uris} \) are the default attributes to appear in the result (See an earlier section (Specifying Relational Attributes to Appear in the Output) ).

There is one point that users have to be careful in using aliases. As explained in an earlier section (Intersection, Union and Difference), the default interpretation of a sequence of subqueries \( (q_1 \ q_2 \ldots) \) is the intersection of the subquery results. However, for subqueries that are related to each other with the same alias, the interpretation is the join of the subqueries. For example, the result of Q15 is the equi-join on \( $a$ \) of the results of \( (\text{class:Film as $a$}).\text{starring.}(\text{class:Actor Tom}) \) (the same as the result of Q10 (Figure 13)) and \( a.\text{starring.}(\text{class:Actor Ken}) \) (Figure 16). The attributes of the result relation of Q15 are \( t\_\text{uris} \) of the two subqueries.

Formal Semantics and Relational Completeness

This section defines the formal semantics of Gradation Language. Then, we show that Gradation is relationally complete. Actually, the expressive power of Gradation is equivalent to the relational algebra with text containment predicate, which means that the queries can be translated into relationally complete structured query languages with text containment predicates (such as SPARQL).

This section is organized as follows: First, we map the data \( D \) (explained in The Data Model) to an equivalent universal relation. Second, we give the formal semantics for queries that we explained (see Gradation Query Language) only intuitively. Finally, we show that Gradation is relationally complete in the sense that it can express any relational queries on the universal relation.

Mapping Text and RDF Data to a Relation

This section explains how to map \( D \) (text and RDF data) to a universal relation \( U \) that preserves the information in \( D \). For the explanation, we use the data shown in Figures 17 and 18.

The data in Figure 17 have two text pages Samurai_t and Ken_t. Figure 19 shows the result of mapping it to \( U \). Let \( D = (T, G, f) \). Then, \( U \) contains one tuple for each resource node \( n_t \) in \( G \). Attributes of \( U \) are \( r\_\text{uri}, t\_\text{uri}, \text{text}, \) and attributes corresponding to all of the predicates that appear in \( G \). (e.g.,
Each tuple contains values of the \( r_{\text{uri}} \) of \( n_t \), \( t_{\text{uri}} \) of \( t \) s.t. \( n_t = f(t) \), and attribute values each of which is \( o \) s.t. \((n_t, p, o) \in G\). The values of attributes that do not correspond to predicates on \( n_t \) are null.

In general, a resource node can have two or more values for one predicate. For example, in Figure 18, Samurai_r has two edges for each of “starring” and “type” predicates. Figure 20 shows the result of mapping it to \( U \). When \( n_t \) has \( m > 1 \) values for the same predicate \( p \), \( U \) contains \( m \) tuples that are the same as each other except that attribute values for \( p \) are different. This recursively applies to the general case in which two or more such predicates exist. For example, \( U \) has four tuples for Samurai_r (Figure 20), which are generated by combining \{Ken_r, Tom_r\} for “starring” predicate and \{Film, HistoricalFilm\} for “type” predicate.

**Constructing the Universal Relation.** This section explains the algorithm for translating \( D \) into \( U \). The algorithm works with the simple case in which each RDF node has only one value for each predicate. The algorithm for the general case in which nodes have two or more values for the same predicate is a natural extension of the algorithm and thus omitted. The algorithm (Figure 21) works as follows. Given \( D \), it generates a tuple for each resource node \( n_t \in N \) (Lines 2-3), stores attribute values in each tuple (Line 4-13), and inserts the tuple into \( U \) (Line 14). Here, node_uri(\( n_t \)) in Figure 21 is a function to get the URI of resource node \( n_t \). Similarly, textpage_uri(\( n_t \)) is a reverse function of \( f \) to get the URI of text page corresponding to \( n_t \) (i.e., \( n_t = f(\text{textpage_uri}(n_t)) \)). text(\( u \)) is a function to get the contents of the text page of URI \( u \).

**Theorem 1** The algorithm in Figure 21 outputs \( U \) that preserves the information stored in \( D = (T, G, f) \). □

**Proof.** To prove the theorem, we have to show \( U \) preserves information stored in \( G \), \( T \), and \( f \), respectively. First, we show that \( U \) preserves the information in \( G \). Lines 10-13 guarantee that for every \( n_t \in N \), there is a tuple in \( U \) that has information on every triple (\( n_t, p, o \)) in \( G \). Each tuple in \( U \) contains node_uri(\( n_t \)) in the attribute \( r_{\text{uri}} \) and contains \( o \) in the attribute \( p \).

Next, we show that \( U \) preserves the information stored in function \( f : T \rightarrow N \). Remember that \( f \) is a one-to-one mapping between text pages and RDF resource nodes and that for every text page there is one RDF node in \( N \). Given the fact, Line 8 guarantees that for every \( n_t \), there is a tuple that has URI of the corresponding text page (if any). Since \( f \) is an injective function and the URI of every \( n_t \) is stored in \( U \), it preserves the information encoded in function \( f \).

Finally, we show that \( U \) preserves the information in text pages in \( T \). Line 9 guarantees that each tuple in \( U \) stores not only the URI of a text page but the contents of it, which means that \( U \) preserves the information stored in the contents of text pages in \( T \). □

**Formal Semantics**

This section gives the formal semantics of queries that we explained only intuitively in an earlier section (Gradation Query Language). First, we give the formal semantics of Gradation queries in terms of relational algebra expressions on the universal relation \( U \). Second, we explain the formal semantics for subtle cases in which the relations are not union compatible in Intersection, Difference or Union operations.

**Semantics of Gradation Queries.** Table 2 summarizes the semantics of Gradation queries in terms of relational algebra expressions on the universal relation \( U \) we explained in an earlier section (Mapping Text
Figure 20. Universal relation $U$ for the data in Figure 18

**INPUT:** Data $D = (T, G, f)$, where $G = (N, P, E)$

**OUTPUT:** Universal Relation $U$

1: $U = \emptyset$
2: FOREACH $n_t \in N$
3: $\text{Tuple } tuple = \text{new Tuple}();$
4: FOREACH $A_i \in U's \text{Attributes}$
5: $tuple[A_i] = \text{NULL};$
6: ENDFOREACH
7: $tuple[r_uri] = \text{node_uri}(n_t);$
8: $tuple[t_uri] = textpage_uri(n_t);$
9: $tuple[text] = text(textpage_uri(n_t));$
10: FOREACH $(n_t, el, v_2) \in E$
11: IF ($v_2$ has URI) $tuple[el] = \text{node_uri}(v_2);$
12: ELSE $tuple[el] = v_2;$
13: ENDFOREACH
14: $U.insert(i);$
15: ENDFOREACH
16: RETURN $U;$

Figure 21. Algorithm to create the relation $U$

and RDF Data to a Relation. In Table 2, $q$ is a Gradation query and $\text{sem}(q)$ is a relational algebra expression that is equivalent to $q$ (modulo the relational attributes that appear in the output). $U$ is the universal relation, and $R.a$ is the attribute $a$ of relation $R$.

We use the table to define the semantics of $q$ as follows: Given $q$, the semantics of $q$ is defined as $\pi_{\text{attrs}}(\text{sem}(q))$, in which $\text{sem}(q)$ is given in Table 2, and $\text{attrs}$ are the set of attributes that are tagged as “to appear” by $q_1[\text{attribute}_1, \ldots, \text{attribute}_n]$ in $q$. Therefore, we first construct a relation $R = \text{sem}(q)$ with Table 2. During the process, we compute $\text{attrs}$ when we encounter $q_1[\text{attribute}_1, \ldots, \text{attribute}_n]$ in $q$. As explained in an earlier section (Specifying Relational Attributes to Appear in the Output), the expression can be nested and $\text{attrs}$ will be overwritten by the outer one. Finally, we project out the attributes that are not contained in $\text{attrs}$ from $R$.

Resolution Rules for Union Incompatibility. This section explains the resolution rules for union incompatibility in Intersection, Difference and Union operations. An informal explanation was already given in an earlier section (Combining Keyword Queries with Path Traversals).

Union. When two relations for the union are not union compatible, the union operation returns the outer union (Navathe & Elmasri, 2002) of the two relations.

Intersection. In general, the two relations $\text{sem}(q_1)$ and $\text{sem}(q_2)$ for the intersection must be union compatible. However, Gradation defines the result of the intersection if one of $q_i$s is a keyword query $k$ (say, $q_1 = k$). In that case, we interpret the $k$ and $q_2$ as selecting tuples in $\text{sem}(q_2)$ that contains $k$ in any of $\text{text}$ attributes of $\text{sem}(q_2)$. Formally, the result of query "$k \text{ and } q_2$" is $R = c_{\text{text}_1 \text{ contains } k \text{ and } \text{text}_n \text{ contains } k}(\text{sem}(q_2))$. Here, $\text{text}_i$ is a $\text{text}$ attribute of $\text{sem}(q_2)$. Note that the intersection is commutative thus "$q_2 \text{ and } k$" returns the same result of "$k \text{ and } q_2$".

Difference. Similarly, the two relations $\text{sem}(q_1)$ and $\text{sem}(q_2)$ for the difference must be union compatible. However, Gradation defines the result of the difference if one of $q_i$ is a keyword query. Formally, the semantics of query "$k - q_2$" is $c_{\text{text}_1 \text{ not contains } k \text{ and } \text{text}_n \text{ not contains } k}(R_k)$. Here, $k$ is a keyword, and $\text{text}_i$ is a $\text{text}$ attribute of $\text{sem}(R_k)$. The semantics of query $q_1 - k$ is $c_{\text{text}_1 \text{ not contains } k \text{ and } \text{text}_n \text{ not contains } k}(R_1)$.
The unique point of Gradation is its syntax - some queries can be written as pure keyword queries, some queries can be equivalent to five relational operations.

### Relational Completeness

This section proves the following theorem.

**Theorem 2** Gradation is relationally complete on the universal relation $U$ shown in an earlier section (Mapping Text and RDF Data to a Relation).

**Proof.** We show that Gradation can express queries that are equivalent to five relational algebra operations: Union, Difference, Cartesian product, Projection and Selection on $U$. In the discussion, we assume that:

1. Relation $R$, $S$, $T$ and $V$ are any relations that can be derived by applying relational algebra expressions to $U$.
2. $R$ and $S$ are union compatible,
3. The schema of $R$ is $R(A_1, A_2, \ldots, A_i, \ldots, A_j, \ldots, A_n)$,
4. $\theta$ is one of $<, \leq, >, \geq, =, \neq$, and
5. $\epsilon$ is a constant.

Then, Figure 22 shows how the five operations can be expressed in Gradation.

### Evaluation

As shown in an earlier section (Formal Semantics and Relational Completeness), the expressive power of Gradation is equivalent to the relational algebra with text containment predicate. The unique point of Gradation is its syntax - some queries can be written as pure keyword queries, some queries can be
written by combining keywords and non-keyword components, and others can be written as pure structured queries, allowing us to write queries in a pay-as-you-go fashion. In order to confirm that Gradation allows us to write many keyword queries and hybrid queries other than pure structured queries, we counted the number of each type of queries that can be written in Gradation.

Query Styles

We define three query styles to which each Gradation query expression belongs. Note that only the syntax matters to define the styles. It is possible that query expressions in different styles are equivalent in their results.

- **Pure Keyword-query Style**: A query in this style has an appearance of Boolean keyword queries. An example is \((\text{Tom or Thomas}) \text{ and actor}\).

- **Pure Structured-query Style**: A query in this style is a direct translation of a relational algebra expression into Gradation. An example is \(\text{Class:Actor}[\text{name}]\), which is a direct translation of \(\pi_{\text{name}}(\sigma_{\text{Type} = "Actor"}(U))\).

- **Hybrid Query Style**: A query in this style can be classified into neither pure keyword nor structured-query styles. An example is \(\text{Tom class:Actor}\).

Experiment

The purpose of the experiment is to confirm that Gradation allows us to write query expressions in various styles so that we can write queries in the pay-as-you-go fashion. In the experiment, we counted the number of Gradation query expressions in each of the three query styles. As explained in Introduction, support to change queries into more precise ones is one of our future work.

In the experiment, if we don’t limit the number of words that appear in a query and the height of the concrete syntax trees of queries, it is obvious that every query style has a countable infinite number of queries and the comparison is meaningless. Therefore, we parameterize the number of words \(|W|\) and the height of the concrete syntax tree (height) in counting the number of queries. Here, \(W\) is a set of words used in queries. In the experiment, we assume that any \(w \in W\) can appear at the places in which keywords, class names, attribute names, or string values can appear in a query.

We compared the numbers of queries in each of the three styles, varying the values of height and \(|W|\).

Results

Figure 23 shows the number of possible query expressions in each query style in the case of \(\text{height} = 8\) and \(1 \leq |W| \leq 50\). Figure 24 shows the number of possible query expressions in the case of \(|W| = 1\) and \(5 \leq \text{height} < 20\). The figures show that Gradation allows us to write many hybrid queries in addition to pure keyword and structured ones (Note that the Y-axis is in log scale).

Conclusion

This paper proposes Gradation, a hybrid query language for text and associated structured data that covers both keyword queries and structured queries. Users can add various query components to keywords.
to constitute various kinds of queries in Gradation. We proved that Gradation is relationally complete and showed that Gradation allows us to write many hybrid queries in addition to pure keyword and structured queries. Therefore, Gradation allows us to write queries in a “Pay-as-you-go” fashion, i.e., describe various types of queries, according to the tradeoff between the acceptable cost of writing queries and the required preciseness of the query description.

Future work includes the development of efficient processing schemes for Gradation queries, the development of a scheme for ranking query results, and the development of a support system to show the user hints so that she can rewrite her imprecise queries (typically keyword queries) to more precise, structured ones.

References


Appendix: The Syntax of Gradation Query Language

<query> ::= <exper>
<exper> ::= <term>
| <exper> ’ or ’ <term>
<term> ::= <factor>
| <term> (‘ and ’ | ‘ ’ | ‘ * ’ | ‘.’ <edge> ‘.’) <factor>
| ‘not ’ <term>
| <term> as <alias>
| <term> <cond>
| <term> ’[’ <attrs> ‘]’
<factor> ::= (‘ <exper> ’)
| <keyword>
| ‘class:’ <classname>
<edge> ::= <exper_edge>
<exper_edge> ::= <term_edge>
| <exper_edge> ’ or ’ <term_edge>
<term_edge> ::= <factor_edge>
| <term_edge> (‘ and ’ | ‘ ’) <factor_edge>
| ‘not ’ <term_edge>
<fact_edge> ::= (‘ <exper_edge> ’)
| <edgename>
<cond> ::= <exper_cond>
<exper_cond> ::= <term_cond>
| <exper_cond> ' or ' <term_cond>
<term_cond> ::= <factor_cond>
| <term_cond> (' and ' | ' ') <factor_cond>
| 'not ' <term_cond>
<fact_cond> ::= '(' <exper_cond> ')'
| <attrname> <relational_op> ( <attrname> | <constant> )
<attrs> ::= <attrs> ',' <attrname>
| <attrname>

<keyword> ::= <string> | <uri>
<constant> ::= <string>
<attrname> ::= '-'? ( <edgename> | 't_uri' | 'r_uri' )
<classname> ::= <identifier>
<edgename> ::= <identifier>
<alias> ::= '$' <identifier>

<relational_op> ::= '>' | '<' | '>=' | '<=' | '=' | '!='
<string> ::= [0-9a-zA-Z\*\-\./]+
| '"' [0-9a-zA-Z\*\-\./][0-9a-zA-Z\*\-\./]* '"'
<identifier> ::= [a-zA-Z][a-zA-Z0-9.]*
A Dyadic Approach to Information Mediation at Work: Examining Credibility and Value Perceptions

Ji Yeon Yang  
School of Information  
University of Michigan  
Ann Arbor, MI 48109-1285  
jlyeon@umich.edu

Soo Young Rieh  
School of Information  
University of Michigan  
Ann Arbor, MI 48109-1285  
rieh@umich.edu

Abstract

In daily interaction, workers play the dual role of information seekers and mediators by receiving or providing advice on how to find and use information. Using an online diary method, this study examines the dynamic and interactive process of information mediation focusing on (1) what factors influence how workers perceive the credibility of advice, (2) what factors influence how they perceive the value of the information mediation process, and (3) how their credibility perception impacts the value perception, depending on whether they receive or provide advice. The results show that, when receiving advice, credibility and value perceptions were almost exclusively influenced by the nature of the task for which the advice was needed. When providing advice, those perceptions were affected by more diverse factors including advice type and tenure. Furthermore, the relationship between credibility and value perceptions showed a marked difference depending on whether a person received or provided advice.

Keywords: workplace information behavior, information mediation, credibility, diary study

Introduction

Seeking information through colleagues is an important part of daily work. Compared with seeking information from document sources, the precision of queries is less crucial in seeking information from colleagues; instead, through conversation, two people can reach a mutual understanding regarding problem definition and what information is needed (Bruce et al., 2003; Robinson, 2010). Several researchers have identified those colleagues as information mediators, who intervene in the information seeking process of others by providing guidance and advice (Ehrlich & Cash, 1994, 1999; Kuhlthau, 2004). The concept of information mediation illuminates how information seeking at work becomes a social process in which strategies for finding and using information are developed and negotiated.

Information mediation is often invisible because of its embeddedness in daily interactions between workers (Ehrlich & Cash, 1999) as they play the dual roles of information seeker and mediator. Workers transition seamlessly between receiving and providing advice in finding or using information. In the process of information mediation, workers who provide advice would transfer their perspectives and judgments of information to colleagues who seek and receive advice, potentially influencing the seekers’ subsequent information behaviors. The information seekers, however, may not accept their colleagues’ advice as it is. Their acceptance depends upon the extent to which they perceive the advice to be credible and find the information mediation process to be valuable.

Taking a dyadic approach to studying information mediation at work, the ultimate purpose of this study is to better understand these dynamics in information mediation from the perspectives of both information seekers and mediators. We believe that the interactive process of information mediation bears further research given the current work environment in which organizations increasingly adopt social media and knowledge management tools. These tools provide workers with diverse communication channels through which they can intervene in each other’s information seeking processes while unknowingly influencing one another.

Specifically, this study examines workers’ credibility perception of advice as well as their value perception of information mediation. Previous studies about information behavior in organizational
settings tend to focus exclusively on the perspectives of information seekers, while paying relatively little attention to the perspectives of information mediators. This study presumes that advice shared in the process of information mediation and the value of the process itself are evaluated not only by those who receive advice but also by those who provide it. Unlike previous studies, therefore, this study examines the dual perspectives of information seekers and information mediators to gain a more complete understanding of trust perception and evaluation in the process of information mediation at work.

This study addresses the following research questions:
1. What factors influence how workers perceive the credibility of advice shared in the process of information mediation?
2. What factors influence how they perceive the value of the information mediation process?
3. How does their perceived credibility of advice relate to their perceived value of information mediation?
4. How do their perceived credibility of advice, value of information mediation, and the relationship between the two differ depending on whether they receive or provide advice?

In order to address these research questions, it is necessary to capture in-the-moment experiences of both advice-receiving and advice-providing in natural settings. We therefore conducted this study within a real-world workplace setting, using a diary method that combines signal- and event-contingent designs (Wheeler & Reis, 1991).

**Literature Review**

Studies have consistently found that organizational workers often rely on their colleagues for information. For instance, Allen (1977) found that engineers and scientists were nearly five times more likely to turn to a person for information than to an impersonal source such as a database. Despite the development of enterprise search systems, people still consider gathering information through personal contact with colleagues critical to the success of projects (Cross, 2000) and consult their colleagues as an “entry point to the written documents” (p. 11). A few studies have identified the significance of interpersonal information seeking in the workplace. According to Zipperer (1993), seeking information from colleagues is beneficial in that (1) they can provide feedback; (2) their memory might be the only way to access a document; and (3) they enable the selection of trustworthy experts within a particular subject domain. Kraut, Fish, Root, and Chalfonte (1990) argued that informal communication is necessary for organizational coordination given the nature of novelty, unexpectedness, and uncertainty in organizations. Organizational workers, therefore, deliberately build, maintain, and activate personal networks (Nardi, Whittaker, & Schwarz, 2000).

These studies indicate that information seekers have demonstrated a need for someone to be there for guidance during the process of finding and using information, as information mediators (Kuhlthau, 2004). While there is limited literature available about information mediators, existing studies have revealed their essential role in the process of seeking and using information. In the context of library service, Kuhlthau viewed information mediation as an intervention in a user’s search process (2004). Her study distinguished between source- and process-oriented mediation, emphasizing the importance of the latter in seeking meaning. By observing customer support organizations, Ehrlich and Cash (1994) found that informal information mediation occurs across the process of daily work, from correctly diagnosing a problem, identifying, evaluating, synthesizing, interpreting, and applying information. The value of information mediation is often invisible to information seekers even if they frequently rely on mediators to identify problems and to learn what kind of information is available (Ehrlich & Cash, 1999). These literature collectively acknowledge the significant yet under-recognized roles of information mediators in the flow of knowledge. To better understand the dynamics of information mediation, however, more research is needed from both the perspectives of the information seeker and the information mediator about how people trust the advice they receive or provide and how they benefit from the process.

In examining people’s perceptions of advice, the credibility research literature provides several attributes that influence whether or not people will believe the information. Credibility has been defined as a combination of trustworthiness and expertise (Hovland, Janis, & Kelley, 1953) that determines the believability of information (Fogg & Tseng, 1999). Fogg and Tseng categorized credibility into four types including presumed, reputed, surface, and experienced credibility. In their framework of credibility assessment, Hilligoss and Rieh (2007) identified three distinct levels of assessment: construct, heuristics, and interaction. Their construct level pertains to how people conceptualize credibility; examples included
truthfulness, believability, trustworthiness, objectivity, and reliability. Credibility studies to date, however, have been conducted primarily in the context of online information seeking mainly from the perspectives of information seekers. Comparatively, for organizational work settings, credibility issues related to the use of organizational information have not been the subject of significant investigation. It is important to expand the scope of research to include issues related to judging the credibility of organizational information, because they are closely related to organizations’ decision-making capacities. The present study aims to fill these gaps.

Methods

Online Diary Surveys

In order to capture events surrounding information mediation in situ, we chose to conduct an online dairy survey. Diaries not only enable participants to use their own words in recording events and feelings (Poppleton, Briner, & Kiefer, 2008), but also help them recall memories of those details during subsequent interviews. To examine how workers perceive credibility and value during information mediation from the perspectives of both information seeker and mediator, we developed two sets of diaries: (1) advice-receiving diaries for recording activities during which participants get advice from their colleagues in seeking or using information and (2) advice-providing diaries for recording activities during which participants give advice to their colleagues in seeking or using information. Participants were signaled via corporate email twice a day, at noon and 4 PM, for two weeks (excluding weekends). Each participant was asked to record advice-receiving diaries for one week and advice-providing diaries for the other week. In order to control any order effect, it was instructed that half of the participants start with advice-receiving diaries while the other half start with advice-providing diaries.

Participants

This study was conducted in the R&D department of a large Midwestern manufacturing company. The department consists of over 500 employees including scientists, technicians, and engineers. Previous studies have found that scientists and engineers tend to be highly motivated and active consumers of information (Fidel & Green, 2004; Hertzum & Pejtersen, 2000). Most of the company’s projects are performed across multiple divisions, bringing together people with diverse backgrounds and expertise. The R&D department was chosen as a research site because, according to an initial interview with a divisional director, information mediation between colleagues is essential and is encouraged to accomplish daily tasks. An email invitation to the study was sent out to the entire R&D department and 86 individuals agreed to participate in the study.

Data Collection

Before collecting the diary data, we administered a background questionnaire that includes basic demographic information, job roles, department, and work tenure. Once participants submitted their background questionnaire, they were entered in the system and received e-mails with links to the online diary survey for the next two weeks.

Both sets of diaries consisted of open-ended, Likert-type, and multiple-choice questions. They first asked participants to think about situations during the past four hours in which they turned to their colleagues (or their colleagues turned to them) for work-related advice or information and to choose the one that took the most time. In the advice-receiving diaries, participants were then asked to report characteristics of the task on which they needed help, urgency and complexity of the task, names of up to five people they turned to, method they used to find and communicate with each person, reason they chose each person, characteristics of the advice received, action taken as a consequence of receiving the advice, credibility of the advice received, and value of the information mediation. In the advice-providing diaries, they were asked to report characteristics of the task on which they provided help, complexity of the task, name of the person they assisted, method used to communicate with the person, reason for why they were chosen, extent of the advice they provided, credibility of the advice they provided, and the value of the information mediation. In both diaries, task complexity was measured by
asking participants to indicate how complicated they felt the task on which they received or provided advice was on a scale of 1-7, with 1 being the least and 7 being the most complicated. Previously, researchers have studied task complexity in relation to uncertainty about or a priori determinability of task outcome and information requirement (Byström & Järvelin, 1995). Task urgency was measured on the 1-7 scale as well, but only in the advice-receiving diaries. This is because the timeframe or priority of a task is known by the person who needs help but not by the one who provides help.

Table 1 shows the measures of advice credibility and the value of information mediation that were used in the diary surveys. In advice-receiving diaries, four measures of advice credibility were developed based on previous credibility literature (Hilligoss & Rieh, 2007; Rieh, Kim, Yang, & St. Jean, 2010) and five measures of the value of information mediation were developed based on Saracevic and Kantor’s (1997) taxonomies of values resulting from information services. Those taxonomies were developed in the context of library services, but provide this study with fundamental insights that serve to distinguish between cognitive, affective, accomplishments, and time values. All measures were rated on 1-7 scales.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Description</th>
<th>Measures</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trustworthy</td>
<td>The extent to which they trusted the advice</td>
<td>Expert</td>
<td>Self-rating of their expertise in the topic on which they provided help</td>
</tr>
<tr>
<td>Reliable</td>
<td>The extent to which they relied on the advice</td>
<td>Trustworthy</td>
<td>The extent to which they trusted the advice they provided</td>
</tr>
<tr>
<td>Valuable</td>
<td>The extent to which they found the advice valuable</td>
<td>Confident</td>
<td>The extent to which they felt confident that their advice would be helpful</td>
</tr>
<tr>
<td>Agreeable</td>
<td>The extent to which they agreed with the advice</td>
<td>Satisfied</td>
<td>The extent to which they felt satisfied with the advice they provided</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accepted</td>
<td>The extent to which they think their advice was accepted</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time well-spent</td>
<td>Level of the feeling of time well-spent after consulting the person</td>
<td>Time well-spent</td>
<td>Level of the feeling of time well-spent after the conversation</td>
</tr>
<tr>
<td>Certain</td>
<td>Level of the feeling of certain after consulting the person</td>
<td>Certain about what I knew</td>
<td>Level of the feeling of certain about what they knew after the conversation</td>
</tr>
<tr>
<td>Learned new things</td>
<td>Level of the feeling of learned something new after consulting the person</td>
<td>Learned new things</td>
<td>Level of the feeling of learned something new after the conversation</td>
</tr>
<tr>
<td>Problem-solved</td>
<td>Level of the feeling of problem-solved after consulting the person</td>
<td>Opinion changed</td>
<td>Level of the feeling of opinion changed after the conversation</td>
</tr>
</tbody>
</table>

Data Analysis

After removing incomplete and inappropriate records, the data set consists of a total of 450 diaries, 206 advice-receiving and 244 advice-providing, submitted by 75 participants. About half of the participants started with advice-receiving diaries \((N=35)\), while the rest started with advice-providing diaries \((N=40)\). On average, each participant submitted 2.8 advice-receiving \((SD=1.8)\) and 3.3 advice-providing \((SD=2.1)\) diaries.

The first step of data analysis was to develop a coding scheme to systematically analyze the two open-ended questions: (1) tasks on which the participants needed or provided advice; (2) advice they received or provided. Table 2 shows main and sub-categories of task types used for content analysis, with examples. The tasks were first categorized into five main types: (1) increase descriptive knowledge; (2) increase procedural knowledge; (3) assess value; (4) determine actions; and (5) obtain data.
Descriptive and procedural knowledge were further categorized into technical and non-technical knowledge. Technical knowledge refers to the knowledge of mechanical or scientific issues; non-technical knowledge includes knowledge of business, cultural, or managerial issues. This categorization is relevant to Taylor’s (1991) categorization of information use situations in that both distinguish between factual, instrumental, and personal or political information. Table 3 shows the main and sub-categories of advice types with examples. Advice was categorized into three main types: (1) knowledge addition; (2) value addition; and (3) alternatives suggestion. Compared to the other advice types, value addition involves information mediators’ judgments and personal opinions to a greater degree with an intention of influencing recipients.

Table 2
Coding Categories for Task Types

<table>
<thead>
<tr>
<th>Task type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase descriptive knowledge</td>
<td>Gain technical know-what Understand the functionality of an ingredient</td>
</tr>
<tr>
<td></td>
<td>Gain non-technical know-what Enhance knowledge on the business part of the company</td>
</tr>
<tr>
<td>Increase procedural knowledge</td>
<td>Gain technical know-how Develop a matrix of tests for a project</td>
</tr>
<tr>
<td></td>
<td>Gain non-technical know-how Understand how to build trust within a team</td>
</tr>
<tr>
<td>Assess value</td>
<td>Evaluate Determine whether a presentation covered the right information</td>
</tr>
<tr>
<td></td>
<td>Verify Double check a test procedure the company uses</td>
</tr>
<tr>
<td>Determine actions</td>
<td>Decide Select which sampling plan is best</td>
</tr>
<tr>
<td></td>
<td>Solve Encounter a problem with a piece of equipment during a test</td>
</tr>
<tr>
<td></td>
<td>Plan Set goals around a future team-building event</td>
</tr>
<tr>
<td>Obtain data</td>
<td>Need a statistical summary of data for a project</td>
</tr>
</tbody>
</table>

Table 3
Coding Categories for Advice Types

<table>
<thead>
<tr>
<th>Advice type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge addition</td>
<td>Aggregation Collect raw data and summarize it</td>
</tr>
<tr>
<td></td>
<td>Background knowledge Go through the background of a previous testing</td>
</tr>
<tr>
<td></td>
<td>Experience sharing Share one's approach to a similar problem reflecting on past situations</td>
</tr>
<tr>
<td></td>
<td>Explanation/demonstration Walk through an example of building a new report</td>
</tr>
<tr>
<td>Value addition</td>
<td>Idea/opinion Review and provide comments on a test analysis</td>
</tr>
<tr>
<td></td>
<td>Suggestion Provide a direction based on original scope of work</td>
</tr>
<tr>
<td></td>
<td>Validation Confirm the agenda for an upcoming meeting</td>
</tr>
<tr>
<td></td>
<td>Solution Identify options to prevent incident from occurring</td>
</tr>
<tr>
<td>Alternatives suggestion</td>
<td>Referral to documents/files Supply documentation of a team's future plan</td>
</tr>
<tr>
<td></td>
<td>Referral to other people Provide the name of a person and coach on how to bring up an issue</td>
</tr>
</tbody>
</table>

In the remainder of this paper, we report our findings from analysis of the diary data, focusing on how different factors, including characteristics of individuals, tasks, and advice, affect the perceived credibility of advice and the perceived value of information mediation, and how credibility and value perceptions are related to each other.

Findings

Characteristics of the Participants

As shown in Table 4, among 75 participants, 37 were male and 38 were female. They were distributed across age groups, with the highest concentration (31%) between 45 and 54. The mean duration of work tenure at the company was 10.9 years (SD=9.28). Job roles were diverse, including scientists, technicians, managers, and project managers. More than half of the participants were scientists, including product developers, sensory scientists, and chemists.
Factors Affecting Perceived Credibility of Advice and Value of Information Mediation

To examine which factors affect the perceived credibility of advice and the perceived value of information mediation in both advice-receiving and advice-providing situations, we analyzed the diary data with a linear mixed model followed by Bonferroni post hoc tests (the results from the post hoc tests are not included in the tables). As the diaries were collected multiple times from the same individuals, the responses are nested within individual participants in both sets of diaries. In advice-receiving diaries, the responses were not only nested within the participants but also within the tasks because participants were asked to report up to five people they turned to for a single task and answered the same set of questions for each person. To account for the possible dependencies of the responses, we used linear mixed model analyses with individual participants as a random effect for the advice-providing diaries and with individual participants and tasks as a random effect for the advice-receiving diaries.

For the advice-receiving diaries, we analyzed the effect of seven different factors on the workers’ credibility and value perceptions. The factors included individual characteristics such as gender and tenure (see Table 4); task characteristics such as task type (see Table 2), task urgency, and task complexity; advice type (see Table 3); and for cases with multiple sources of advice, the order in which advice was received. Table 5 shows the F statistics of the effects for each of those factors on nine credibility and value measures, controlling for the other covariates in the model.

Table 4
Demographics of the Participants

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Category</th>
<th>Frequency (n=75)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>18-24</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>25-34</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>35-44</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>45-54</td>
<td>23</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>55-64</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>65+</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>38</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>37</td>
<td>49</td>
</tr>
<tr>
<td>Job title</td>
<td>Scientist</td>
<td>38</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Technician</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Manager</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Project Manager</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Regulatory/legal</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Administrative</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Engineer</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Job tenure</td>
<td>&lt;1year</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1~5 years</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>5~10 years</td>
<td>20</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>10~20 years</td>
<td>22</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>20~30 years</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>&gt;30 years</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 5  
Advice-receiving Episodes: F Statistics of the Effects for Factors on Credibility and Value Measures

<table>
<thead>
<tr>
<th>Credibility of advice received</th>
<th>Gender</th>
<th>Tenure</th>
<th>Task type</th>
<th>Task urgency</th>
<th>Task complexity</th>
<th>Advice type</th>
<th>Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trustworthy</td>
<td>1.74</td>
<td>1.26</td>
<td>2.79**</td>
<td>5.00*</td>
<td>0.37</td>
<td>0.94</td>
<td>3.21</td>
</tr>
<tr>
<td>Reliable</td>
<td>0.52</td>
<td>0.87</td>
<td>1.49</td>
<td>6.78*</td>
<td>1.57</td>
<td>0.62</td>
<td>2.60</td>
</tr>
<tr>
<td>Valuable</td>
<td>3.17</td>
<td>0.82</td>
<td>2.62**</td>
<td>4.13*</td>
<td>1.21</td>
<td>0.72</td>
<td>1.17</td>
</tr>
<tr>
<td>Agreeable</td>
<td>0.92</td>
<td>2.08</td>
<td>2.88**</td>
<td>4.93*</td>
<td>4.30*</td>
<td>1.31</td>
<td>2.74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value of information mediation</th>
<th>Time well-spent</th>
<th>Time certain</th>
<th>Time satisfied</th>
<th>Time problem-solved</th>
<th>Time learned new things</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.04</td>
<td>0.24</td>
<td>0.04</td>
<td>0.02</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>0.97</td>
<td>0.24</td>
<td>1.12</td>
<td>2.27</td>
<td>2.47*</td>
</tr>
<tr>
<td></td>
<td>2.94**</td>
<td>3.06**</td>
<td>2.99**</td>
<td>2.04*</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>3.92*</td>
<td>7.29*</td>
<td>3.59</td>
<td>11.53**</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>1.78</td>
<td>9.09*</td>
<td>8.59**</td>
<td>33.94**</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>1.91</td>
<td>1.91</td>
<td>1.26</td>
<td>1.49</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>2.01</td>
<td>0.53</td>
<td>3.40</td>
<td>0.39</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Note: *p<.05, **p<.01

Overall, the perceived credibility of the advice received and the perceived value of information mediation were influenced almost exclusively by the nature of tasks - task type, task urgency, and task complexity. On the other hand, individual characteristics such as gender and tenure had no significant effect on credibility and value perceptions, excepting a slight difference in perception of learned new things depending on tenure. The lowest Bonferroni adjusted p-value (p=0.11) was found between those who worked less than a year (estimated marginal mean=6.32, SE=0.42) and those who worked more than 20 years and less than 30 years (estimated marginal mean=4.68, SE=0.42). Advice type, such as whether the advice added knowledge, added value, or suggested alternatives, as well as the order that the advice was received had no significant effect on credibility and value perceptions.

We now look more closely at the effect of the nature of the task on credibility and value perceptions. First, task type was a significant predictor of several perceptions, including trustworthy (p=0.01), valuable (p=0.01), agreeable (p=0.00), time well-spent (p=0.00), certain (p=0.00), satisfied (p=0.00), and problem-solved (p=0.04). We performed post-hoc tests to determine for which task types the workers found advice less credible or information mediation less valuable. Across all measures except problem-solved, verify was the only task type that was significantly different from at least one other task type on the 0.05 level. This indicates that, when advice was received on the task of verifying, the workers tended to find the advice less trustworthy, valuable, and agreeable, and felt the time less well-spent, less certain, and less satisfied compared to when advice was received on other tasks. For problem-solved, there were no significant differences found among different task types after adjusting for multiple comparisons. The lowest Bonferroni adjusted p-value (p=0.29) was found between gain non-technical know-what (estimated marginal mean=6.02, SE=0.40) and solve (estimated marginal mean=4.61, SE=0.32).

Second, analysis reveals that task urgency had a significant effect on trustworthy (p=0.03), reliable (p=0.01), valuable (p=0.04), agreeable (p=0.03), time well-spent (p=0.05), certain (p=0.01), and problem-solved (p=0.00). The coefficient estimates show the positive association between task urgency and these six credibility and value measures. The largest coefficient estimate (0.27) was observed in the effect on problem-solved, which indicates that the perceived level of task urgency led to the greatest positive change in the average for the feeling of problem-solved after receiving advice.

Lastly, task complexity had a significant effect on agreeable (p=0.04), certain (p=0.00), satisfied (p=0.00), and problem-solved (p=0.00). The coefficient estimates show a negative association between the task complexity and these four credibility and value measures. The largest coefficient estimate (0.46) was observed in the effect on problem-solved, which indicates that the perceived level of task complexity leads to the greatest negative change in the average for the feeling of problem-solved after receiving advice.

For the advice-providing diaries, we analyzed the effect of five different factors on the workers’ credibility and value perceptions. The factors included individual characteristics such as gender and tenure (see Table 4); task characteristics such as task type (see Table 2) and task complexity; and advice...
When providing advice, the workers’ perceived credibility of their advice and their perceived value of information mediation were influenced by more diverse factors such as tenure, nature of task, and advice type, compared to when receiving advice.

We now look more closely at the effect of each of those factors. First, job tenure was a significant predictor of the perception of expert \( (p=0.02) \) and trustworthy \( (p=0.03) \). For the perception of expert, post-hoc tests revealed that those who worked less than a year rated their expertise in the topic on which they provided advice significantly lower \( \text{(estimated marginal mean}=4.79, SE=0.39) \) than those who worked more than 10 but less than 20 years \( \text{(estimated marginal mean}=6.02, SE=0.18) \), those who worked more than 20 but less than 30 years \( \text{(estimated marginal mean}=6.49, SE=0.25) \), and those who worked more than 30 years \( \text{(estimated marginal mean}=6.47, SE=0.38) \) on the 0.05 level. For the perception of trustworthy, post-hoc tests revealed that those who worked less than a year trust the advice they provided significantly less \( \text{(estimated marginal mean}=5.44, SE=0.31) \) than those who worked more than 30 years \( \text{(estimated marginal mean}=6.75, SE=0.33) \) on the 0.05 level.

Second, task type was another significant predictor of the perception of satisfied \( (p=0.01) \) and time well-spent \( (p=0.01) \). We performed post-hoc tests to examine for which task types the workers found their advice less satisfied and the information mediation less time well spent. Interestingly, for both measures, evaluate was the only task type that was significantly different from at least one other task type on the 0.05 level. This indicates that the workers tended to be less satisfied with the advice they provided and found the time less well spent when they provided advice on the task of evaluating than when they provided advice on other tasks.

Third, task complexity had a significant effect on confident \( (p=0.02) \), satisfied \( (p=0.01) \), accepted \( (p=0.00) \), learned new things \( (p=0.00) \), and opinion changed \( (p=0.04) \). Coefficient estimates show a mix of positive and negative association between task complexity and those credibility and value measures. Task complexity was positively related to learned new things \( (0.45) \) and opinion changed \( (0.15) \), but was negatively related to confident \( (-0.09) \), satisfied \( (-0.09) \), and accepted \( (-0.13) \). This indicates that perceived task complexity led to the greatest positive change in the average for the feeling of learned new things after providing advice, while it led to the greatest negative change in the average for the perception of how well their advice was accepted.

Lastly, advice type had a significant effect on the perception of confident \( (p=0.03) \). Post-hoc testing, however, showed that there are no significant differences between different advice types on the 0.05 level, after adjusting for multiple comparisons. The lowest Bonferroni adjusted p-value \( (p=0.14) \) was found between explanation/demonstration \( \text{(estimated marginal mean}=6.47, SE=0.17) \) and experience sharing \( \text{(estimated marginal mean}=5.64, SE=0.27) \).
Relationship between Perceptions of the Credibility of Advice and the Value of Information Mediation

To examine how perception of the credibility of advice influences perception of the value of information mediation, we analyzed the relationship between individual credibility and value measures in both sets of the diaries. We used the same linear mixed model as in the previous section to control for the possible dependencies of repeated measures data. Prior to analysis, all ratings were standardized to aid in interpretation. Table 7 shows the standardized coefficient estimates between credibility and value measures in advice-receiving diaries.

Table 7
Advice-receiving Episodes: Standardized Coefficient Estimates between Credibility and Value Measures

<table>
<thead>
<tr>
<th>Value of information mediation</th>
<th>Credibility of advice received</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trustworthy</td>
<td>Reliable</td>
<td>Valuable</td>
<td>Agreeable</td>
</tr>
<tr>
<td>Time well-spent</td>
<td>0.61**</td>
<td>0.50**</td>
<td>0.62**</td>
<td>0.56**</td>
</tr>
<tr>
<td>Certain</td>
<td>0.46**</td>
<td>0.45**</td>
<td>0.50**</td>
<td>0.45**</td>
</tr>
<tr>
<td>Satisfied</td>
<td>0.55**</td>
<td>0.48**</td>
<td>0.57**</td>
<td>0.58**</td>
</tr>
<tr>
<td>Problem-solved</td>
<td>0.28**</td>
<td>0.22**</td>
<td>0.32**</td>
<td>0.30**</td>
</tr>
<tr>
<td>Learned new things</td>
<td>0.36**</td>
<td>0.30**</td>
<td>0.46**</td>
<td>0.37**</td>
</tr>
</tbody>
</table>

Note: * p< .05, **p<.01

Looking at the impact of credibility perception on value perception, all four credibility measures significantly impacted all five value measures. When we compared the magnitude of those standardized coefficient estimates, credibility measures had a slightly bigger impact on *time well-spent* and *satisfied* than on the rest of the value measures. Among the credibility measures, *agreeable* had the greatest impact on *satisfied* (standardized coefficient estimates: 0.58, p=0.00). This suggests that when the workers receive advice, their feeling of satisfaction is more strongly influenced by how much they agreed with the advice than by how much they trust, rely on, or value the advice.

Table 8 shows the standardized coefficient estimates between credibility and value measures in advice-providing diaries.

Table 8
Advice-providing Episodes: Standardized Coefficient Estimates between Credibility and Value Measures

<table>
<thead>
<tr>
<th>Value of information mediation</th>
<th>Credibility of advice provided</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expert</td>
<td>Trustworthy</td>
<td>Confident</td>
<td>Satisfied</td>
<td>Accepted</td>
</tr>
<tr>
<td>Time well-spent</td>
<td>0.12</td>
<td>0.32**</td>
<td>0.34**</td>
<td>0.31**</td>
<td>0.54**</td>
</tr>
<tr>
<td>Certain about what I knew</td>
<td>0.47**</td>
<td>0.47**</td>
<td>0.44**</td>
<td>0.50**</td>
<td>0.28**</td>
</tr>
<tr>
<td>Learned new things</td>
<td>-0.04</td>
<td>-0.02</td>
<td>0.05</td>
<td>0.03</td>
<td>0.13</td>
</tr>
<tr>
<td>Opinion changed</td>
<td>-0.1</td>
<td>-0.06</td>
<td>0.02</td>
<td>-0.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note: * p<.05, **p<.01

Looking at the impact of credibility perception on value perception, none of the credibility measures significantly impacted *learned new things* and *opinion changed*. However, the credibility measures had a significant impact on *time well-spent* and *certain about what I knew*, except that self-perceived *expertise* had no significant impact on the feeling of *time well-spent* after providing the advice. Among the credibility measures, *accepted* had the greatest impact on *time well-spent* (standardized coefficient estimates: 0.54, p=0.00). This finding indicates that after the workers provide advice, their feeling of time well-spent is most strongly influenced by their perception of how well their advice was accepted than how much they trusted, felt confident about, or were satisfied with their advice.
Discussion

The results of this study indicate that the workers perceive the credibility of advice shared in the process of information mediation and the value of the process differently depending on whether they receive or provide advice. When receiving advice, the workers’ credibility and value perceptions were mainly influenced by the nature of tasks such as task type, task urgency, and task complexity. When providing advice, their credibility and value perceptions were influenced less by task characteristics than when receiving advice. Rather, more diverse factors including tenure and advice type affected their credibility perceptions. Previous research has focused primarily on the perspectives of information seekers while neglecting the dyadic relationship between those seeking and those providing information. The findings of this study suggest the importance of exploring more factors related to individual characteristics and the nature of advice in understanding the perspectives of mediators who provide information.

A particularly interesting finding was for which task type the workers tended to perceive advice as less credible or place a lower value on the information mediation. When receiving advice, the workers found advice related to the task of verifying significantly less trustworthy, valuable, and agreeable, and felt less certain, less satisfied, and that the time was less well-spent. A possible explanation is that the need for verification arises from discrepancies between their previously existing knowledge and information at hand. Consequently, even after consulting colleagues for advice, they may have lingering uncertainty which makes them trust the advice less and find the process less beneficial. On the other hand, when providing advice, the workers found advice on the task of evaluating significantly less satisfying and felt the time was less well-spent. A previous analysis on task complexity (Yang & Rieh, 2012) showed that the workers perceived the task of evaluation as most complicated when providing advice, but as least complicated when receiving advice. This indicates that the complexity of value judgment has a negative effect on self-perception of success in assisting colleagues.

Another interesting finding was that perceived task complexity produced some positive values when advice was being provided. That is, the more complicated the workers perceived a task to be, the more likely they felt that they learned new things after providing advice. This is contradictory to advice-receiving situations, in which perceived task complexity was negatively associated with the feeling that the problem had been solved. This resonates with previous research demonstrating that task complexity increases information seekers’ needs for problem-solving information such as the methods of problem treatment (Byström & Järvelin, 1995). Interestingly, the perceived level of task urgency was positively associated with the feeling of problem-solved, which indicates the significant effect of time pressure on judgment of the advice-receiving experience.

This study also attempts to identify how credibility perception of advice affects value perception of information mediation from the perspectives of both those receiving advice and those providing advice. When receiving advice, the workers’ satisfaction with information mediation was most strongly dependent on how much they agreed with the advice rather than on how much they trusted, valued, or relied on the advice. This reflects people’s preference for hearing points of view in agreement with what they already understand or believe. When providing advice, the workers’ perception of time well-spent was most strongly dependent on how well their advice was accepted. This indicates that information mediators find the advice-providing experience more rewarding in the presence of positive feedback or reaction to their advice, and supports Lin’s (2007) research about motivational forces in organizational knowledge sharing. According to Lin, intrinsic motivation such as reciprocal benefits, self-efficacy, and enjoyment in helping others is more crucial in sharing information than external organizational rewards. As organizations implement social software to increase inter-organizational knowledge sharing, it is essential to provide recipients of advice with effective ways to explicitly show their appreciation for the advice provided.

232
Conclusion

Based on the diary survey collected for two weeks, this study investigated information mediation in the workplace from both information seekers’ and information mediators’ perspectives. Our results revealed the complexity and dynamics of seeking and providing advice in organizational work settings. In terms of theoretical contributions, the findings indicate that it is important to investigate the judgment of credibility and evaluation of value based on both information seekers’ and information mediators’ daily work practices, because they could report different experiences of information mediation. By analyzing the data captured in-situ in various information mediation contexts, we were able to identify that the seekers’ credibility and value perceptions were mainly influenced by the task characteristics, while the mediators’ perceptions were affected by more diverse factors, such as advice type and work tenure. The findings of this study have practical implications for designing and implementing social media and other knowledge management tools in the workplace. To facilitate the process of information mediation, the system needs to support information activities and to keep tracking information mediation experiences not only for seeking advice but also for providing advice. In addition, the categories of task and advice type should be incorporated into the system design as those factors influence the credibility and value perceptions. For example, for the task types, such as evaluating, for which the mediators feel less satisfied with their own advice, the system must ensure that they can efficiently and securely refer the seekers to other people with relevant expertise for further guidance.

More research is needed to better understand the outcome and implications of information mediation in the workplace. As a follow-up study, we conducted in-depth interviews with 45 diary survey participants to further investigate how the workers enter into the information mediation process and how they influence one another’s subsequent information behavior and decision-making. The interviews elicited rich descriptions of interpersonal trust between information seekers and mediators, assessments of trust in the advice shared, and the effect of the mediation process on subsequent behaviors. The future analysis of interview data should be able to provide deeper insights into the trust and influence between information seekers and mediators.

References


Acknowledgements: This project was funded by the IEEE Robotics and Automation Society (RAS). We would also thank our collaborator Dr. Peter Asaro for his input in the formulation and realization of the project.


Copyright is held by the authors.
development of an interdisciplinary field of research along with cross-cultural differences, scholarly communication, and collaborative practices in science.

In the sections below, we describe the conceptual foundation and rationale for creating digital archive that will address three primary audiences: the robotics community aiming to inscribe and distribute a history of their field, scholars of science studies who want to understand scientific development, and public users seeking information and using the archive for educational purposes. By analyzing and presenting multiple sources of data through the prism of the “ecologies of knowledge” (EOK) framework, we seek to enable the development of rich descriptions of how robotics has developed as a global and interdisciplinary scientific field over the past fifty years, analyze changes within and across various social, epistemic, and temporal units of analysis, and bridge local and systemic methodologies for studying science. We aim to create rich networks connecting individuals, events, materials, concepts, and institutions relying on the roboticists’ own life narratives and the narratives that can be surmised from their publication records. Furthermore, we suggest that using new information technology can help us produce a non-linear digital archive that will enable these different audiences to construct and represent multiple stories of robotics and roboticists. We also aim to enhance our understanding of different forms of “memory practices” in the sciences (Bowker, 2005) by comparatively analyzing our two main sources of data—oral history interviews and journal articles—as well as by tracing the activities of abovementioned different types of users as they navigate the archive.

Representing robotics as a knowledge ecology

The development of science has been studied using a span of theoretical frameworks focusing on different levels of analysis: individual careers (Keller, 1983; Conway & Siegelman, 2005), situated lab groups (Latour & Woolgar, 1979; Knorr-Cetina, 1999), disciplines and fields (Gibbons et al, 1994), and the systemic dynamics of science (Merton, 1973; Gieryn, 1999; Leydesdorff & Etzkowitz, 2001). As our archive draws primarily on information from personal oral history narratives and field-wide bibliometric data and seeks to represent a polydirectional view of scientific development, we chose to conceptualize robotics as an “ecology of knowledge” (Star, 1995; Akera, 2007b). This framework allows us to situate and represent a wide variety of human and nonhuman actors evolving through mutual associations across multiple social, epistemic, and temporal units of organization: individual careers, groups with shared worldviews and practices, institutions and their history, and scientific concepts and areas of study. We also used this framework as a guide when developing interview questions and choosing topics for bibliometric analysis, and as the organizing logic behind our database and visualizations.

Cast as an “ecology of knowledge,” science is an activity performed in a social context to which “lived, experienced differences as embodied in specific locales and moments, and communities, are central” (Star, 1995, p. 3). This ecological framing of knowledge production is complemented by an evolutionary understanding of scientific development, which focuses on tracing the relationships between individual and systemic, cognitive and cultural change (Wimsatt & Griesemer, 2007). Development and change in the ecology occurs across different social units of analysis—the individual, artifacts, concepts, groups, institutions, practices, events, cultures (Akera, 2007a)—and across varying temporal scales—lifetimes and habits, generational concepts and practices, and cultures, which define semiotic meanings and group affiliations (Caporael, 2007).

Akera’s (2007a, 2007b) depiction of knowledge ecologies as a series of interrelationships within and between the institutional, social, cognitive, historical, and material factors that play a role in the development of technoscience suggests a non-linear view of history, with multiple possible starting and ending points, paths of development, relationships of influence, and communities involved. In the case of our project, envisioning science as a knowledge ecology allows us to organize and mobilize the information we receive from individual scientists about important events, artifacts, personal relationships, organizations, and other experiences in the context of results obtained from bibliometric analyses of their published works. We represent both a local and personal understanding of robotics and a broader systemic picture that displays “the linkages between knowledge and its various contexts” (Akera, 2007b, 413).
Memory practices in robotics

This project uses two main data sources: oral history narratives collected by interviewing roboticists and bibliographic data consisting of publication records, patents, and information from curriculum vitae. These data sources can all be viewed as memory practices, defined by Bowker (2005) as “acts of committing to record (such as writing a scientific paper)” (p. 7), which are deeply embedded in scientific practices. Different genres of memory practices, such as personal narratives and publications, provide different contexts, audiences, and purposes and can in turn produce quite different descriptions of the past. These practices also often involve the use of various information technologies, which affect both the ways and meanings of producing records of scientific practice.

The project being described was started by members of IEEE RAS who, concerned that information about early developments in the field may be lost to posterity due to the lack of proper documentation and the passing of individuals involved in the field’s early days, suggested performing interviews with early participants in the field. We started with an initial list of “pioneers”—individuals identified by members of the robotics community as prominent researchers—and built up a broader interviewee list through snowballing. Interview subjects are distributed among governmental, industrial and academic institutions in the United States, Europe and Japan. While most existing archives, such as the Charles Babbage Institute and the IEEE History Center’s Oral History Collection, present written transcripts of interviews, we decided the focus of our archive will be on presenting thematically related short video clips from the original interviews to enable more intuitive access to the lived experiences of robotics researchers, while also allowing access to transcripts when specifically desired by the users. We currently have 100 interviews and over 150 hours of video recorded, and are continuing with data collection.

Influenced by the “ecologies of knowledge” (EOK) framework, our interview questions focus on learning about how particular individuals got involved in robotics; what their aims were in doing robotics research; what were the important institutions and events that affected their work; how their work developed through the years; who they were inspired by; who they had connections with in the robotics industry and outside of it; what kinds of social and cultural factors affected their work, etc. The interviews allow us to develop a view of robotics from the perspective of its practitioners, to identify the key individuals, relevant institutions, ideas, and events that have played a significant part in their careers, and provide us with data not available in written records. Such data is particularly useful in revealing how individual values and actions have shaped robotics, and how specific past experiences are shaping present-day actions and values.

As a memory record oral histories provide description of the past from a scientist’s perspective. In that respect, they are rather selective in terms of scientists having a fairly limited space and time to portray their career and their role in the development of a particular research area. They present an opportunity for a researcher to shape the future views of himself/herself and the whole discipline they belong to. They also provide an opportunity to champion a particular worldview and reconfigure the past. In addition, oral history interviews are grounded in the memories of individuals, which are a subjective and partial reconstruction of events in the past.

The figure below shows an “ecology of knowledge” representation developed using one of our interviews with Professor Ruzena Bajcsy (UC Berkeley). The layers of the ecology are populated with people, artifacts, events, occupations, organizations, and institutions mentioned by Prof. Bajcsy during the interview, relationships among the different components of the ecology are presented using lines and

1 Interviewees include individuals at Berkeley, Stanford, JPL, NASA, Precise Automation, Adept Technologies, Willow Garage, UPenn, CMU, MIT, USC, Unimatic, Johns Hopkins, iRobot, Caltech and Georgia Tech in the US; University of Tokyo, Tokyo Metropolitan University, Tsukuba University, AIST, Kyoto University, Nagoya University, ATR, Kawada Industries, Osaka University, Honda, Waseda University in Japan; Seoul National University and Kwangwoon University in South Korea; LAAS Toulouse, EPFL Switzerland, ETH Zurich, TU Munich, University of Karlsruhe, KUKA Robotics, ABB, CNR Rome, Oxford, Univ. of Hertfordshire, Edinburgh, KU Luewen, and KTH Sweden in Europe.

2 http://www.cbi.umn.edu/

follow the narrative presented by the interviewee. As a personal narrative, the number of different actors and concepts mentioned is smaller than what we obtain from analyzing bibliographic data related to Bajcsy's work, but it brings out which aspects of her experience are emphasized in her own self-narrative, and describes the interconnections among people, materials, ideas, historical events, and organizations in ways that are not possible (or would be very difficult) using bibliometric methods. Particularly interesting are Bajcsy's descriptions of the relationship between technology and her work, as well as the development of interdisciplinary collaborations with biophysicists, archeologists, and dancers.

**Figure 1.** Presented here is the ecology of knowledge representation of one of our interviews. The objects are placed on eight different planes, with full lines showing relations within a plane, and dashed lines connections between planes. Following one thread of symbolic and practical relationships, marked in red, we can start with the link between communist philosophy and the need for more electrical engineers, electrical engineering (EE) as a discipline and Slovak Technical University (STU), where Bajcsy got her MS and PhD in EE with a thesis relating to machine learning. From there we are connected with Prof. Gvozdiak, Bajcsy's advisor at STU who met John McCarthy from Stanford, where Bajcsy went to get her PhD in Computer Science. Following from that is an assistant professorship at Penn State and connections with robotics, various artifacts (PUMA Arm, French Finger), and multiple collaborators.
Along with oral history interviews, we chose to use bibliographic data, the second major data source for this project, to develop a systemic study of robotics and help contextualize the individual standpoints of pioneers in the larger social and epistemic structures in which they emerge. We are creating a digital library containing a collection of documents related to various topics in robotics relying on bibliographic information, including titles, abstracts, publication years, author names, associated organizations, references and citations for journal articles, conference proceedings, technical report, dissertations and patents.

Journal articles in particular have an exceptional communication and archival function in science. Because of their role as a repository of research results, journal articles make an excellent data source. Bowker (2005) considers journal articles “the central medium for the dissemination and exchange of scientific ideas” (p. 126). Journals also serve as a way of establishing disciplinary boundaries; in that respect they can be treated as institutions. As a memory practice the primary purpose of journal articles is to persuade the intended audience of worth, originality, and cognitive authority. Through the careful usage of language and references to support the findings, scientists use this type of memory practice primarily to persuade their contemporaries of the quality and worthiness of their research. Along with serving the scientists in a particular field to display and disseminate their knowledge, journal articles have also become the main data under analysis for a large number of systemic studies of science.

We treat journal articles not only as a particular genre of memory practice, but also as a complementary source of data that we can use to situate and further elaborate on some of the aspects of roboticists’ oral history narratives. Below we show two examples of possible uses of bibliometric analysis within the archive and our studies of scientific developments in robotics. Figure 2 displays Ruzena Bajcsy’s collaborators and includes a much larger group of people than we were able to discuss in the interview. It also gives us a quantitative measure of the strength of Bajcsy’s ties to particular collaborators, which may not be the same as her own understanding of saliency in the interview. Indeed, the strongest co-authorship connection between Bajcsy and her student T. Sobh did not come up during our interview. We are also able to see some of the interconnections among her collaborators, which seem particularly strong among the individuals in the upper right hand cluster that includes her advisor, J. McCarthy. Figure 3 is a “heatmap” of 50 most frequently occurring words used by the robotics community in the titles of their published work. Heatmaps allow us to visualize trends in the usage of terms across article publication years. It suggests that there are four phases in the development of robotics: 1983-1986, 1987-1994, 1995-2010, and 2011- onwards. From this visualization, we can surmise that Bajcsy’s early interest in computer vision was indeed pioneering, as the term is constantly in use but really gains in popularity in the last few years.
Figure 2. Collaboration network for Ruzena Bajcsy. The network shows 107 co-authors on the papers in which R. Bajcsy was the first or corresponding author.
Figure 3. Heatmap of the relative frequencies of the 50 most frequently occurring terms across the time period from 1983 to 2011.
Aside from journal articles, twentieth century science is remarkable for the significant increase in the contributions of different forms of scientific publications (such as conference proceedings, standards and patents) that have not been so prominent at earlier times. The development of industry in 19th century Europe and the United States opened up new opportunities for scientists to work in research and development (R&D) laboratories, so that they no longer had to be exclusively tied to universities and research institutes (Vickery, 2000). Industry not only provided jobs to some scientists, but started supporting research at universities by endowments, which in turn raised issues of technology development and transfer. While this has previously been resolved by a physical presence at a site of new technology development, the rate of technological development precluded the efficiency of this approach. Patents and patent laws appeared as a good solution to this problem. A patent can be defined as “a legal instrument which gives a temporary monopoly to an inventor in exchange for detailed publication of the invention” (Nesta & Patel, 2004) (p. 533). Thus a patent serves two purposes: protecting the inventor while allowing him/her to make a profit from their invention and enabling wide dissemination of knowledge about the invention. According to Vickery (2000) “by 1990, about three-quarters of a million patents were being issued annually throughout the world, and the total volume of patents exceeded 27 million” (p. 149). However, patents have not been equally used by scientists and engineers in different types of institutions. Meadows (1998) claims “10-20% of academic scientists and engineers use patents frequently, as compared with some 80% in industry” (p. 224). According to Walker (1995) industry is not only a major consumer, but also a major producer of patents, with about 80% of patents being held by mainly large corporations. Patents are thus very useful for studying the aspects of research at the borderline of science and technology. In our project patents primarily inform us about the developments in industrial and technical aspects of robotics, which have had an important effect on the development of robotics and serve as a memory practice related to industry, about which it is more difficult to obtain data through publications and publically available resources.

Similarly to the results from bibliometric analysis, patent analysis can show that interviewees may have different priorities and viewpoints of their achievements than is suggested by publically available data. In interviews, George Devol, one of the creators of the Unimate robot that IEEE RAS cites as the beginning of “robotics history,” describes robotics as one among many interests and fields in which he has made innovations (Ballard et al, 2012). The patent record, however, shows that his robotics-related patents are among the most influential and most often cited by others in the field (See Figure 4.)

---

**Figure 4.** Deicted above are patents granted to Devol from 1938-1984, along with their relative frequency of citation by other patents expressed by the height of the lines. Only patents that have at least 10 citations are depicted. Following the late 50s, you can see numerous robotics-related patents.

---

4 The early system of patents appeared in Venice in 1474. The first patent in England dates from 1552; American patent law was passed in 1790 and French in 1791 (Vickery, 2000).
All of the above data sources have been extensively used to study scientists and their disciplinary lineage. They have also been used to understand how scientists themselves have been making sense of the entities they have been studying. However, these different sources have been used by different people, in different contexts and often provided incommensurable stories of science and scientists. We are interested in not only using these records to tell stories of robotics, but to understand the meaning and place of these memory practices in the careers of scientists and practice of science. The combination of these data sources and analyses will be used not only to develop a more comprehensive understanding of the evolution of robotics, but to evaluate the utility of combining so called situated and systemic approaches to science studies.

Combining situated and systemic approaches to understanding science

In order to understand the evolution of a knowledge ecology, we must analyze both the individual level of scientists’ experiences as well as the species level of the field as a whole. This poses the methodological challenge of bridging interpretive analysis of individual narratives, which provide a rich interpretative understanding of local interactions among scientists, artifacts, and environments, with systemic analyses of large data sets that give a broad view of the structures and dynamics of knowledge production.

Local, situated studies of science identify aspects of scientific practice and culture by analyzing the interactions among scientists, institutions, and technologies in local environments of knowledge production, “in modern science, typically the scientific laboratory” (Knor-Cetina, 1995, p. 140). Such studies rely on interviews and observation to understand scientists’ personal experiences, discursive strategies, and the everyday practices and dynamics of science. Laboratory studies and ethnographies mark a move in science studies from the study of science as knowledge towards the consideration of “science as practice” - what scientists do, along with the context and product of their work (Pickering 1992, p. 2). Oral histories also seek to describe science from the perspective of its participants, focusing on meaning-making, the interactions between people, concepts, and material artifacts, and the influence of different cultural and social factors on the process and results of scientific work. While situated approaches provide researchers with “thick descriptions” (Geertz, 1973) of local interactions and practices, they are hard to generalize and replicate due to unique nature of the contexts they study. Scientometric approaches, in contrast, provide a systemic perspective using information theoretic tools to analyze documents and other textual artifacts (e.g., articles, grants, patents), map scientific communication, and understand how population-level structures in science develop over time and across communities. Most of these types of studies focus on mapping science (visualization), developing indicators to be used in science policy, and more recently studying science as social networks. Although networks, a set of nodes and a set of ties between those nodes, have played a major role in studies of science both as a metaphor and as a method (e.g., Ben-David & Collins, 1966; Crane, 1969, 1972; Mullins, 1972, 1973), it is only with the advancement of computing technology that large-scale network studies of disciplines have become possible. Numerous studies by physicists use coauthorship networks to study network dynamics (Barabasi & Albert, 1999; Barabasi et al., 2002; Farkas et al., 2002; Newman, 2001). One attempt to use large-scale networks to study the structure of a scientific discipline from a sociological standpoint was done by Moody (2004), who studied the structure of social science collaboration networks by connecting network topologies to empirical and theoretical findings from the sociology of science. While the network approach is successful at describing structures and their change over time, it does not represent the situated processes leading to specific social and epistemic structures, or the individual’s perspectives and meanings regarding science.

This project attempts to bridge these two approaches to studying science primarily by exploring novel ways of combining usage of a wide variety of data sources through a nonlinear digital archive of robotics.

A non-linear archive supporting diverse memory-making practices

To represent the kind of heterogeneous and multidirectional relationships suggested by Akera’s depictions of ecologies of knowledge, we have had to consider what the most fruitful way to present all these data is. We find this aspect of the project very important since the technology and representation
we choose will influence not only the usage of the archive, but the types of stories one can tell about the robotics. Our decisions in that respect are guided by the envisioned usages. The archive should allow roboticists to explore their past and learn more about their discipline. At the same time we envision the archive as a powerful tool that researchers who study science can use to study the development of disciplines; collaboration; cross-cultural differences within science; the co-construction of science, technology and society; and finally nature of different memory practices within science and their role in communication and archiving of scientific knowledge. The archive can also be used by educators and the general public to construct their own histories and stories of robotics and roboticists.

Current on-line oral history archives are often forced into flat linear structures. Such archives in turn compel users to follow a singular story line constructed by the creators of the archive. We want to take advantage of full capabilities of current technology to allow for non-linear presentations of narratives and data that do not conform to rigid timelines nor are forced into presenting a single aspect of the phenomenon. Thus, by “non-linear” we mean the ability of viewers to choose to continue watching a video, or to follow a link and move fluidly to watching other meaningfully related videos, as one can choose to follow hypertext links in webpage text. The ability to follow non-linear narratives is especially important in oral history, where one can examine an issue, topic, relationship, or event from multiple perspectives.

We will use recent innovations in HTML5 to allow interaction with edited versions of our oral history videos, as well as select network visualizations, analysis results, and other documents, available online. The videos will be edited according to the thematic coding of the interviews, designating particular components of the knowledge ecology—people, institutions, instruments, theories, ideas, concepts, etc.—or certain themes that emerge from data analysis. Elements of the ecology of knowledge identified in oral histories and coded themes will be matched and linked to similar documents existing in the bibliographic data. This will allow for direct comparison of social, epistemic, and institutional networks created from these two sources of data.

Conclusion

This paper provides conceptual foundation for the creation of a non-linear digital archive of oral histories of roboticists. The archive is based on the conceptualization of science as an “ecology of knowledge”, composed of diverse sets of human and nonhuman actors evolving through associations across multiple social, epistemic, and temporal units of organization: individual careers, groups with shared worldviews and practices, institutions and their history, and scientific concepts and areas of study. The paper discusses the roles of different data sources as memory practices primarily used in situated and systemic studies of science. It is through this different memory practices that different communities interested in robotics negotiate, construct and represent their narratives of robotics and roboticists. The paper also discusses the role that this novel archive can have in the process of reconfiguration of past and present of the field and collective memory practice building.

References


Towards a Data Literate Citizenry

Michael B. Twidale  
twidale@illinois.edu

Catherine Blake  
cblake@illinois.edu

Jon Gant  
jongant@illinois.edu

Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign

Abstract

We believe that data literacy should be a skill not just for scientists, but for all citizens. We make the case by considering literatures on various kinds of literacy, and using a number of examples to explore some of the challenges that emerge from trying to move to a more data literate citizenry. We consider the opportunities arising from developing technologies to help individual communities and intermediaries in understanding how data should inform personal and societal decision-making.

Keywords: data literacy, citizen participation, open data, information reuse

Introduction

Continuing growth of hardware, software and networking makes it possible to collect and use ever greater amounts of data. Along with opportunities, concerns arise about what we as a society want to allow, encourage or forbid, particularly with respect to privacy, security and ownership of different kinds of data. Debates on these societal and policy issues are themselves informed by analyses of data, as are other debates on priorities, finance, resource allocation, healthcare, climate change, etc.

This current data environment creates a set of concerns over the skills necessary to make intelligent use of available data. What needs to be known and who needs to know it? We see the growth of various professions of data analysts with specialist skills in creating, maintaining, using and interpreting data. However we claim that in a democracy it is important that we try to enable all people to have some level of data literacy in order to be able to fully participate in a discussion around important decisions that a society has to make – informed by ever more data. This is an argument for data literacy based on the importance of traditional literacy (what we might call ‘book-literacy’) in conventional societal participation. As the interest in Big Data continues, we believe that the question will recur of who gets to participate in the discussion and who has to simply surrender to trusting the judgment of experts.

Our goal in this paper is to explore what it means to have a data literate citizenry, and why it matters. This vision leads to a research agenda that builds on earlier ideas of information literacy, but considers the complex information ecosystem where people, information, data, methods to process data, and mechanisms to share all interact. As such, data literacy needs to be considered from multiple perspectives and involve researchers from a range of backgrounds – not just domain experts and curriculum designers, but also systems developers, designers of information visualizations, and researchers in online communities, communication, community informatics, the public understanding of science, informatics and library and information science. Information professionals will remain critical players as teachers, facilitators and intermediaries in data literacy interactions.

Acknowledgement: This paper is made possible in part by a grant from the U.S. Institute of Museum and Library Services (IMLS), Laura Bush 21st Century Librarian Program Grant Number RE-05-12-0054-12.


Copyright is held by the authors.
A Literature of Literacies

Building on established work on regular literacy (the ability to read and write text), researchers have explored other literacies. Even when talking about text, literacy researchers note that the basic ability to read and write is not enough – we should also think about abilities to understand and use what is read and written.

As a result there are now literatures on numerous literacies, each with multiple, often somewhat contradictory and overlapping definitions. For example, there are literatures on literacy, digital literacy, information literacy, scientific literacy, and data literacy. For our focus on data use we also need to consider statistical literacy and indeed basic numeracy. In many contexts, a degree of comfort with using computational technologies is also necessary – something addressed in the computer literacy literature.

In his review article on digital literacy Bawden (2001) notes that in the context of basic literacy (reading and writing), the concept of levels of literacy has proven useful. In highly developed countries few people are completely illiterate, but various studies have shown worrying proportions of the population to be functionally illiterate. Categories such as very low literacy and low literacy are useful in understanding the level of skill needed to make use of textual information in various ways. For example a person with low literacy may be able to read simple text, but be unable to look something up in an alphabetized directory. Bawden clarifies the relationships between various literacies, including a useful history of the development of the idea of information literacy. A later work (Bawden, 2008) provides additional contextualization of the concept of digital literacy and various meanings that different authors apply to different literacies.

Of these literacy terms, information literacy is the one that seems to generate the most interest in the field of Library and Information Science. The Association of College and Research Libraries (ACRL) (2000) defines information literacy as: a set of abilities requiring individuals to “recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information.” This is then elaborated into six elements:

- Determine the extent of information needed
- Access the needed information effectively and efficiently
- Evaluate information and its sources critically
- Incorporate selected information into one’s knowledge base
- Use information effectively to accomplish a specific purpose
- Understand the economic, legal, and social issues surrounding the use of information, and access and use information ethically and legally

Substantial work has explored ways of teaching information literacy skills to all ages; from elementary school, through college and to adults in public libraries. The Big6 approach developed by Eisenberg & Berkowitz (2011) is a particularly well-regarded example.

The importance of information literacy for civic participation has been emphasized for a long time, for example by Owens (1976): “Information literacy is needed to guarantee the survival of democratic institutions. All men are created equal, but voters with information resources are in a position to make more intelligent decisions than citizens who are information illiterates. The application of information resources to the process of decision-making to fulfill civic responsibilities is a vital necessity.” Work in statistical literacy also emphasizes the importance for citizens to understand the issues underlying society and the economy (Podehl, 2003).

Similarly, Shapiro & Hughes (1996) make the case for information literacy as a prerequisite for participation in an information society. They build on the 18th century vision of Condorcet, who advocated for access to information (and the skills necessary to read and use it) so that educated citizens “will no longer depend for every trivial piece of business, every insignificant matter of instruction on clever men who rule over them in virtue of their necessary superiority.” We aim to extend this argument to ideas around data literacy.

In a similar vein, Zuccala (2009) shows how open access to scientific literature can and should benefit not just fellow scientists but also the layperson. Healthcare information is a primary area of interest for many people (particularly that relating to an individual’s or family member’s medical condition), but it is certainly not the only area. Zuccala’s use of the resonant term “laity” reminds us of the risks of delegating all power of use and interpretation to its antonym: priesthood. It is possible that a priestly caste of scientists are the only ones able to collect, analyze, and interpret data, and then remain the only conduit to the laity about what the findings mean for making important societal decisions in lawcourts and
legislative assemblies. With open access to research papers (Zuccala’s point) and underlying data (our extension) the laity can have at least an opportunity to try and verify certain claims and even maybe make their own discoveries. Such freedom entails numerous risks (as do all freedoms). Knowing how to search, select and interpret findings are critical skills to avoid misinterpretation. Access is necessary, but not sufficient. Zuccala’s analysis shows how we can understand current and future needs for civic scientific information literacy by weaving together work on information seeking, science communication in society, and the public understanding of science. This leads to a consideration of the potential contributions of intermediaries including science journalists, museums, scientist-popularizers, and (we claim) librarians and other information professionals.

Schield (2004) connects three literacies (data, statistical and information literacy) all necessary to enable critical thinking skills including analysis, interpretation and evaluation – concepts that we further explore in the next section.

Researchers using the term data literacy have looked at its use in various contexts including enabling educators to use student performance data to improve teaching (Love 2004), and for undergraduates studying geography (Hunt, 2004). Qin & D’ignazio (2010) consider the requirements for a new course addressing data-related literacy for science students. Their investigations revealed the central role that metadata should play in their course. Carlson et al. (2011) use the compound term data information literacy to refer to the skills needed to prepare students to participate in e-research. They used the ACRL six element framework as a way to examine the topics they had derived from an analysis of faculty or student practices and needs, finding considerable conceptual overlap.

Surveying multiple overlapping fields can be rather confusing. There is something of a terminological mess as different authors can mean slightly different things when they use the same x-literacy term, and very similar subconcepts can be part of more than one x-literacy. One way to help make sense of the richness and variety of literacies literature can be to consider issues of scale, genre and use. Different skills are needed at different scales. To cope with, or make use of (that is, read) one book you need basic literacy (book-literacy). To cope with or make use of 100 or 1,000 books (a small library) you also need certain kinds of information literacy. Making the transition from a small school or public library to that of a large university research library (1,000,000 or 10,000,000 books) requires additional information literacy skills. Having access to online journals and indeed the entire web (1 to 100 billion documents) requires yet more information literacy skills. Each of these levels of skills supplement those of the smaller scales – they certainly do not replace them.

As we switch genre from text documents (books, articles, web pages) to data elements (in millions of datasets, each potentially containing billions of values, plus associated metadata of varying form, format, quality and completeness), effective use also requires data literacy. Similarly, reading a single book on science may require a certain amount of scientific literacy (Hazen & Trefil, 2009), basic numeracy, and perhaps statistical literacy. Making use of larger datasets and findings derived from multiple, perhaps heterogeneous datasets requires further skills relating to statistics, estimation, error, and as Qin & D’ignazio (2010) reveal, metadata.

In data literacy just as other literacies we are concerned with not just consumption (reading) but also production (writing). Levels of sophistication matter; much can be improved without everyone becoming an expert. Fluency means an ability to gain an overview as well as follow details. In the context of very large amounts of resources, different access and evaluation skills emerge and grow in importance. Finally, literacies are not just individualistic. They are about participating in a community, and sociotechnical systems can be designed to help that community nurture ongoing learning and growth.

**Motivating Example: Data in the Courts**

As a very simple, constrained example consider the case of statistical evidence in court proceedings. Expert witnesses are brought in, but their testimony needs to be understood by the lawyers, the judge and of course the jury. Problems arise when the level of understanding and (it must be admitted) the clarity of the evidence is lacking. A recent UK case highlights the problem. A 2010 UK Court of Appeal case (Regina v. T.) involved the use of Bayesian methods to assess the rarity of the print from the shoes of the defendant that matched those found at the scene of the crime (Saini, 2011). Under imperfect information about the exact number of particular brands sold in the UK, the expert witness had had to make some informed estimates. The process was perhaps poorly explained and the judge consequently quashed the conviction. Additionally the judge also ruled against using similar statistical
analysis in the courts in future. This has been misreported as an exaggeration claiming that the judge banned all use of Bayesian statistics in UK courts. Rather the judge said that Bayesian methods were an inadmissible way to present expert evidence — except for DNA and “possibly other areas where there is a firm statistical base”. The sales estimates used as a component of the shoe match calculation were considered not “firm”, but of course this opens up a mess of what is to be determined to count as firm.

The case has raised considerable concern. An article in Nature (Fenton, 2011) notes that fallacies of statistical reasoning have influenced verdicts in dozens of widely documented cases. The article reports the creation of: “an international consortium of statisticians, forensic scientists and academic and practicing lawyers (80 people signed up in the first 2 months) to develop guidelines for when and how Bayesian reasoning should be used to present evidence”.

The point of the example is not to point out innumeracy in the British Judiciary (British judges already have a reputation with being out of touch with much popular culture). Rather we want to make the case that as data becomes more widespread, it can and will be used in evidence. This evidence will need to be weighed and will no doubt be contested by the other side in the case. If judges, lawyers and juries don’t understand how to handle evidence derived from data, this is as problematic as a legal system where textual evidence is used but some of those involved cannot read.

Data Literacy for Greater Civic Participation

Data produced and collected by government are the basic ingredients for governments to provide services, make policy, and be held accountable for their performance (Heeks, 1999). “In democratic societies citizens have a basic right to know, to speak out, and to be informed about what the government is doing and why and to debate it.” (Stiglitz, 1999 p.29). Strategies to promote more open and participatory approaches for government center on “using information and communication technologies to operate transparently, facilitate easy and low cost access to public records, and to make civic and social data available in standardized formats that support productive public use of data” (Knight Commission, 2009).

Openness and transparency are being enhanced as governments implement open government initiatives to increase transparency, participation, and collaboration. Open government refers to “government that co-innovates with everyone, especially citizens, shares resources that were previously closely guarded; harnesses the power of mass collaboration, drives transparency throughout its operations, and behaves not as isolated department of jurisdiction, but as something new, a truly integrated and networked organization” (Lathrop & Ruma, 2010). A commitment to access is laudable, but without the appropriate data literacy skills, co-innovation with citizens is unlikely.

A Vision and How to Get There

We envision a sociotechnical ecology where data, information, people and technology co-evolve. That means that we are not advocating for a particular curriculum, but rather for how we might combine resources in different ways to increase levels of data literacy. There are likely to be multiple solutions and indeed different kinds of research that can contribute to these goals. We illustrate those below.

Extending Data Literacy to Citizens

Most of the existing literature and discussion focusses on scientists, trainee scientists, information professionals, and other experts. We want to extend this to considering the general public as a whole, as a data literate citizenry. Other points on this dimension may be non-experts or people with different amounts of expertise but with a particular need. For example how do you get a jury up to speed with data-based evidence, and how is it best presented and discussed?

Similarly, although we increasingly talk about “citizen scientists”, the nature and depth of participation in the scientific process can be very variable. While it is entirely appropriate to have certain activities that are very basic but nonetheless useful to science and meaningful to the participants, it would be good to also provide structures for ever greater participation at greater depth. A useful analogy would be how Wikipedia provides ways for more sophisticated participation not just in editing web pages but adjudicating disputes and managing the entire process. In a similar way we might explore ways that we
can help people migrate from "citizen lab assistants" to something more like true "citizen scientists". Figure 1 best characterizes our vision to include citizens with little or no statistical training into the fold of data literacy.

| Statisticians | Scientists | Citizens |

Figure 1. Work on cyberinfrastructure pushes data literacy from the realm of statisticians to scientists. We envision citizen users with little or no scientific training.

Using Existing Information Infrastructures for Learning

The work of Qin & D’Ignazio (2010) and Carlson et al. (2011) are examples of approaches geared to teaching data information literacy in formal learning contexts (courses offered at universities). There are alternatives of more informal learning, including science museums, book popularizations of science, TV shows, podcasts, etc. There are also approaches that are somewhere inbetween or happen in an overlapping context. For example, information literacy may be taught in a formal lecture, or may arise in a help-giving interaction with a reference librarian that can involve teaching by stealth. In an online health community there may be peer support and activity that is more about incidental or co-learning - where the help giver(s) did not know the approach at the outset, but discovered it along with the help seeker(s). Variation and innovation can occur along this spectrum. In addition to peer-to-peer learning, the growth of MOOCs shows one possible way of enabling very large groups of people to learn certain topics at low cost (Hyman, 2012).

Empowering New Teachers

Online communities around topics such as personal healthcare (Preece 2000) and tech support (Singh & Twidale, 2008) often involve teaching skills of data use and data manipulation. This kind of teaching or help-giving may be done by those formally designated with that task, or as a more spontaneous emergent activity.

If we wish to extend this, how do we design/build/engineer an online community that helps people explore and understand a particular kind of data (say, personal health information)? Or do we just hope that it emerges? What can be built so that more established members can help newcomers to understand a dataset, what to do with it, and how to avoid getting confused or making incorrect inferences? Most online communities are text forums, for obvious reasons of convenience and familiarity. But when talking about data and its creation, interpretation and use, is text really the only and best medium for discussing and explaining? How can visualizations and software use be better folded into these discussions?

Empowering Old intermediaries

There are many kinds of intermediary who may help in supporting greater data literacy; just as different information professionals have been involved with basic literacy and with information literacy. For data literacy these intermediaries include librarians, data curators, journalists, expert witnesses, and scientists acting as popularizers via books, TV shows or online. An example of the latter in the context of data analytics is Hans Rosling (2007), who uses Gapminder data visualizations to help make complex (and often counter-intuitive) points about social development in an accessible way. There is always the risk that intermediaries end up doing the work for the end users, rather than helping the end users do it for themselves. This would be the difference between say a professional letter writer providing a vital service for an illiterate person, and a teacher helping that same person learn to write their own letter. As the analogy indicates, learning data literacy is likely to be time consuming and difficult. It is understandable why people may be daunted, give up or delegate the understanding to experts. As
advocates of a data literate citizenry, we will need to explore how to lower the learning barriers of cost, time and effort.

**Building New Technology Support**

Data literacy discussions often focus on who is teaching whom and what is being taught. This is entirely appropriate. But it is also useful to consider the ways that different technologies can help that process by making it easier to understand and use data. Typically these were initially developed for the benefit of experts, but with some extra work they can be made more accessible to a wider population. This is what happened with library catalogues and many bibliographic databases. Initially they were developed for use by highly skilled librarians who learned complex query languages in order to maximize recall and precision in a single (extremely expensive) query. These became online public access catalogues to libraries with easier to understand and use interfaces intended for patrons. Similar routes were followed by various databases, and of course by search engines. We can aim for similar approaches to data analytic tools.

Examples of tools that could be developed or refined to be more accessible to a general audience include bibliometric search, data mining, text mining, data visualization, claims analysis, and tools to help with understanding of terminology or methods (which may be as simple as Wikipedia lookup and YouTube videos). In addition, a less data literate population is likely to encounter difficulties and need help in how to use tools for different purposes. Consequently there is also a need for technologies to support the discussion of, learning and sharing of skills in online and face-to-face communities.

Some tools may be so powerful as to work in a standalone context for many users (as OPACs often do). An example may be a tool that aggregates and summarizes a literature on a topic in medical research, as well as providing explanations of what it did, and what various technical terms mean. Although it is highly desirable to have standalone applications that can explain something without needing an intermediary, much can still be gained by developing applications that partially help the end user, but still need an intermediary to explain the subtleties.

Tools can be useful for facilitating data literacy even if they are not complete standalone infrastructures that enable all kinds of data analysis or create a “citizen data analytics environment”. It can be useful to have a diversity of tools of varying sophistication, which individuals and groups can assemble to meet particular needs. For citizens this may include many more simple, basic and low cost tools, while scientists require more complex tools that are harder to learn and are more expensive. That is, we need to support ways of combining tools like Many Eyes and Google Refine as well as similar more powerful applications. It remains an ongoing challenge to design tools to optimize various ad hoc assembly and repurposing. Equally, we need to design these tools to enable the learning of data analysis skills and not to be so daunting to learn and use that they serve to further alienate people already somewhat intimidated by datasets and lacking confidence in basic mathematical skills.

**Levels of Data Literacy Sophistication**

Inspired by research in basic literacy highlighting the importance of considering levels of literacy, we think it will be fruitful to develop a better sense of levels of data literacy that permit particular kinds of understanding, interpretation and data use.

For example, a basic level of understanding of a straightforward dataset might involve understanding simple measures (such as mean, median and variance) and how they can be used to assess patterns and trends over time. More advanced concepts may be around how large datasets can be used to predict likely future outcomes, such as the probabilities of different events. Yet more advanced would be ideas like Type I and Type II errors, and simple Bayesian statistics, as in our court case example.

Other cases might be more about a degree of guided literacy – where citizens may not be able to operate independently, but can with help from others; who may not necessarily be experts. For example, we might assess whether citizens can follow a data-driven argument and question things that seem odd. In the context of data driven findings from experts, can we get to the state of the Reagan quote of “trust but verify”? That is, where individuals or groups can (with help) actually check some of what the experts are saying.
In the legal example of the problems with Bayesian reasoning we can see how that might play out. It would certainly be very nice if in the future all school leavers had a rudimentary understanding of these issues, thereby being able to serve effectively as jury members; assessing statistical evidence and able to participate in political debates where data is marshaled and contested. But lobbying to add yet more content to a curriculum is a long term endeavor. In many countries various unintended consequences of high stakes testing make it difficult to focus on more subtle analytic skills as opposed to easily memorizable facts and exam-passing tricks. It would also be nice and somewhat easier to lobby for all people graduating from university to have a basic set of data literacy skills, perhaps as part of a General Education requirement. We would definitely support that, and with greater hope of change happening faster than at the school level. But in the meantime there are ameliorations that are well worth doing.

It is incumbent on experts to acquire not just domain skills but communication skills. For example, what are the best ways of communicating Bayesian statistics so that judges, lawyers and juries can understand them? What are some classic recurrent misconceptions? How might intermediaries develop repertoires of detecting and addressing these misconceptions? How might information professionals help in not just accessing data but co-interpreting it? We believe there is a vital, indeed radical role for librarians, amongst others, in this space.

**Integrating Different Approaches to Data Literacy**

Experts in curriculum design can clearly contribute by identifying core skills and prerequisites. We have seen substantial work in this area in various x-literacies. But we want to emphasize that it is not the only kind of research contribution. More analytic work can help to understand the barriers to data literacy, ranging from better understandings of math and computer phobia, to identifying and understanding commonly recurring misconceptions about data, probability and statistics. Work on misconceptions in physics (Brown, 1992) and statistical reasoning (Tversky & Kahneman, 1974) indicate the potential of misconception analysis.

Builders of technology can explore ways to make data analysis easier to do – or easier to explain to less expert individuals. Improved visualization of results is highly desirable, but there is also a need for better visualizations of other analytic work: how the results were obtained, errors in the results, the claims and counter claims in the literature and step-by-step how-to guides of using tools. Similarly, carefully crafted data visualizations can help in illustrating concepts that we know that people find difficult like Type I and Type II errors.

In the past the barrier was just getting access to the data. This is becoming easier. It is necessary but not sufficient for a citizenry to be in a position to be able to verify claims. As well as understanding the claims of scientists as they relate to policy, data literate citizens can also make use of that data in their lives and in participating in science. We have advocated for data literacy as desirable for civic participation in general, but this can seem a rather abstract idea. A more immediate application of data literacy skills through activities such participating in citizen science, or in personal health management can be a powerful motivator for developing those skills. We now give two scenarios that illustrate what this might look like in the near future.

**Situating Text in the Data Literacy Conversation**

Scientists already use published peer-reviewed literature to inform public health policy. For example, systematic reviews conducted by the Cochrane Collaboration and Health Technology Assessment play an important role in evidence based medicine which in turn can influence government policies on standards of health care. In epidemiology meta-analytic results (a quantitative form of a systematic review) can influence legislation, product label requirements, and public services.

The systematic review process is typically a group activity whereby scientists identify a comprehensive collection of articles, extract information from those articles, verify the accuracy of those extracted facts, and analyze the extracted facts using either qualitative or quantitative techniques (Blake & Pratt, 2006). Although systematic reviews accurately capture evidence, the process is time-consuming, taking 28 months from the original conception through to publication (Petrosino, 1999) and 1139 hours (Allen & Olkin, 1999). With more than 21 million citations in MEDLINE and an additional 1900 new citations added every week, the manual techniques currently used are becoming increasingly difficult to
apply. Consider a breast cancer expert. It would be difficult, but necessary for her to consider the 33,883 articles published on breast cancer during the 28 months required to conduct a systematic review. Faced with the daunting task of sifting through currently available and recently added articles, our breast cancer expert may turn to other strategies to reduce the number of articles, such as constraining her hypothesis or her selection criterion. However, both of these constraints introduce undesirable biases, and thus reduce the validity of her review to inform public policy. Citizen scientists could play an important role in this process by participating in the information extraction activities, for example you don’t have to be expert in medicine to identify the number of people in a study.

The second key challenge of a systematic review is that articles considered are drawn from published literature and thus may suffer from publication bias; where articles that find statistically significant findings are more likely to be published than articles that do not show statistical significance, even though the methodology of both studies are the same. "For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results" (Rosenthal, 1979).

The Multi-User Extraction for Information Synthesis (METIS) system is an example of using text for analytics in that the system automatically identifies information required in a meta-analysis from full text scientific articles. Such an approach can be used within the existing systematic review process to reduce the time between when findings are published and when policy is updated. The automatically extracted information, after being verified by citizens, could be used to create a synthetic control group estimate so that information from articles that would not be considered in a traditional meta-analysis could be incorporated into the analysis. Although this is a controversial suggestion from a meta-analytic perspective, the goal here is not to replace meta-analysis, but rather provide alternative ways to leverage textual "big data".

\[ \text{Breast Cancer and Ever Drinking Estimated Effect Size} = 0.066 \]

Cases 0.517 versus Controls 0.451 (28 studies, n=47158)

**Figure 2.** The METIS Summary of breast cancer articles published between 1997 - 2002.
Figure 2 shows this new type of analysis using breast cancer articles in key epidemiology journals, which were published between 1997 and 2002. Articles that report alcohol consumption as primary information (in the title, keywords or abstract text) are shown as black and articles that report consumption only within the full text (secondary information) are shown in grey. Although scientists who conduct systematic reviews place a high priority on obtaining the “file drawer” articles, studies of the manual systematic review process (Blake & Pratt, 2006) show that users review the abstracts before retrieving the full text, so it is unlikely that these studies would have been found. In one traditional meta-analysis of breast cancer and alcohol consumption (Ellison et al., 2001) 71 of the 72 articles included have “alcohol” or a synonym in the title, keywords or abstract. Some would argue that methodologically studies that do not report the disease and the risk factor as primary information should not be included in a meta-analysis.

The results of this study show that more than 60% of the breast cancer articles that report alcohol consumption (17 out of the 28 studies) include the data as secondary information (only in the full text) and thus would not be included in a traditional meta-analysis. More important than the number of primary versus secondary articles is the degree to which the findings reported in those articles differ. Of the four articles that would be included in a traditional analysis (circled), three suggest that ever drinking is higher in subjects with breast cancer (the cases) and only one study suggests that alcohol consumption is not a breast cancer risk factor. In contrast to the 17 studies that report alcohol consumption as secondary information, 6 show positive effect, 6 show a negative affect and 5 show no effect. Despite including both primary and secondary information, the METIS results are consistent with the earlier cited traditional meta-analysis, which suggests a small positive effect size between ever drinking and breast cancer.

Text is often missing from existing conversations about data analytics, but a manual version of the proposed strategy which took approximately six weeks to complete on 1008 articles [Personal communication T.Tengs, 2002], has been used to quantify the association between smoking and impotence (Tengs & Osgood, 2001). Assuming that a similar amount of time would be required for the 240,000 breast cancer articles such an analysis would take 27 years. Although issues such as access to full text need to be resolved before this approach can be applied to all breast cancer articles, METIS can be used on the subset of articles that are available electronically. Moreover, METIS can reduce the time required for a traditional analysis by automatically identifying information from the articles and thus reduce the time to integrate new scientific findings into public policy.

Making it Personal

People will have more and more access to data about themselves – recording their own activities and with a move to personalized medical care. Looking for health information online is already widespread. As healthcare costs continue to soar a data literate citizenry could help control the expense of managing chronic diseases by accessing the latest scientific information and by incorporating those findings into their daily lives. We begin by describing Morgan, our data literate citizen of the future, as a data consumer.

Morgan remembers a radio story that suggested certain mushrooms could help reduce risk of breast cancer, but she doesn’t know which type of mushroom would be best for her. She uses her computer to identify mushroom study results that are most relevant her stored genetic data. Unfortunately none of her genomic characteristics match the study profiles in the mushroom study, so the system accesses the latest literature on mushrooms. As with any disease or treatment the scientific literature reports many different studies from different countries that each report slightly different results. The system weights each study result based on the quality of the study design and the similarity between Morgan and the subjects used in the study. Finally the system converts the outcomes to more accessible language and visualizations that Morgan can understand and suggests three different types of mushrooms would be beneficial.

Greater access to scientific data in an understandable form is a wonderful first step, but the real benefit of the envisioned data ecology is that Morgan can also actively participate. Let us consider also Morgan as a data producer and analyst.

Based on the system recommendation, Morgan decided to include shiitake mushrooms into her diet. But rather than exploring this alone, she participates in an ongoing community dietary study. Like her neighbors she has also noticed that her cholesterol levels have been lowered, but there seems to be a difference depending on where the mushrooms were purchased. She tells the system to record the date,
time and location of each purchase and the system establishes that particular farms appear to have bigger impacts. Combining these results from other growers around the world reveals that soils with certain trace elements are particularly beneficial.

Conclusion
As well as being essential for doing data driven science, some level of data literacy is becoming increasingly important for lay people to participate fully in society in various ways including democratic decision-making. This is likely to come about through a mixture of technological resources and people: teachers, experts, helpers and peers in communities both online and face to face.

This vision for a data literate citizenry will certainly have some skeptics. Some will argue that a basic understanding of numeracy is critical to use the quantitative data that is available. We don’t dispute this claim, but we do suggest that citizens have a range of online learning resources that were previously unavailable. Moreover, there are people in society (including librarians and amateur help-givers) who want to help others understand. In the current online world, this desire emerges in forums and blogs. Despite evidence that many people have poor numeracy skills, there are still domains where many people have a great interest in statistical data, such for sporting statistics (in various cultures, baseball and cricket seem to be particularly data heavy). Participatory activities such as citizen science and personalized health data collection and analysis as part of a health oriented online community can simultaneously serve as both an impetus to begin acquiring data literacy skills and a means for learning them in a more participatory manner.

Regardless of variable data literacy, we as citizens and as a society already make decisions based on data. The motivating legal example hinges on the necessary uncertainties and the degree to which imperfect data quantified as a range of uncertainties can be factored in with other evidence to determine innocence or guilt. Similarly, our decisions about what to eat, how much to exercise and which medications to take directly affect our very existence. We posit that a data literate citizenry would enable us all to have a higher quality of life. Condorcet’s advocacy for access to information and the means to make use of it as an 18th century Enlightenment value also applies to the skills to interpret 21st century data.

We need to think carefully about exactly which skills and to what level of sophistication people may need them for various purposes. But as with basic literacy, leaving the laity out of the process of access and interpretation creates a degree of dependency on a priesthood of experts. Equally we must acknowledge that there is a very real risk of lay people misunderstanding and misapplying data. But there is no point trying to prevent greater access – it will happen inexorably anyway. What is needed is a way to help far more people make the transitions from data, through information and knowledge, and so to wisdom.

References
Cooperative Visualization: A Design Case

Nathan Prestopnik
Syracuse University
napresto@syr.edu

Abstract

In this design case, the design of visualization tools to support the Laser Interferometer Gravitational Wave Observatory (LIGO) is described. A participatory approach to visualizing a complex computational pipeline was adopted, with the goal of exploring what benefits might be derived when groups of people visualize complex information for themselves. Direct participation in LIGO activities helped IV researchers to combine explicitly codified data from the LIGO computational pipeline with structural knowledge tacitly held by project scientists. Both sources of information were critical to producing meaningful visualizations and progressing design and research efforts. Ultimately, this design case demonstrates how cooperatively creating visualizations can enhance understanding and support group activities and goals. It is a call for more human-centeredness within the visualization literature.

Keywords: information visualization, interaction, design, cooperative work

Introduction

In the current body of information visualization (IV) literature, there are two broad emphases: representation and interaction (Yi et al., 2007). Representation research explores ways to visually represent information on display devices, including algorithms, techniques, and technologies (e.g. S. Card, K. et al., 1991; Feiner & Clifford, 1990; Fekete & Plaisant, 2002; Furnas, 1986; Mackinlay et al., 1991; Robertson et al., 1991). Interaction research studies the dialog that occurs between users of a visual information system and the system itself. Interactions might include filtering data to order it in various ways, drilling down through a display to different levels of detail, zooming, panning, or otherwise manipulating the visual display to achieve the view or perspective that the user is interested in. Of the two emphases, interaction and representation, representation has received by far the most scholarly attention from IV researchers (Chen, 2005; Chen & Czerwinski, 2000; Ellis & Dix, 2006; Thomas & Cook, 2005; Tory & Möller, 2004; Yi et al., 2007), but interaction is of growing interest in the IV community (Chen, 2005). The focus of this present research is on interaction.

Visualization interactions can be split into three categories: representational interactions, cognitive interactions, and creative interactions. Representational interactions include the many types already mentioned; they provide the user opportunities to modify the visual display and organization of information on the screen. Cognitive interactions, described through a variety of models and theories of visualization (e.g. S. K. Card et al., 1999; Chen, 2003, 2006; Spence, 2001, 2007; Ware, 2004), are the purely mental activities that a user will undertake when working with a visualization, setting aside any affordances the visualization artifact itself might have for manipulating the representation of data.

Creative interactions are defined by human involvement in the creative activity of generating visualizations, with the term “creative” indicating the act of creation – the transformation of raw information into a visual representation. Creative interactions can be highly complex and require a great deal of cognitive effort (for example, manually transforming qualitative or quantitative information into something that can be visualized). Creative interactions are less well studied than other forms of interaction, but they are important. Interactions that result in the formation of a visualization will necessarily require individuals to cognitively and representationally interact with raw information and the visualization itself. The same is
true of groups of individuals who cooperate together to produce visualizations, but with additional possible benefits, including the use of collaboratively developed visualizations to foster discussion and achieve group goals.

Broadly, this present research explores the following: What do groups of people gain by visualizing information for themselves? There are many possibilities, including cognitive or analytical benefits, the ability to cooperate and share information more effectively, and the ability to draw new connections or see new patterns within one or more data sets. Using participatory research techniques, these possibilities are explored in a design case: the documented effort to visualize a complex computational pipeline used by astronomers for signal/noise processing of gravitational wave detector data. The different visualizations described in this design case were produced under the aegis of ongoing development of an information retrieval (IR) system to be used by gravitational wave physicists as part of their work. Key users of the visualizations included the astronomers involved in the Laser Interferometer Gravitational Wave Observatory (LIGO) collaboration as well as the IR researchers responsible for producing the retrieval system for this collaboration and the IV researcher (the author) tasked with studying various use scenarios and creating the visualizations themselves.

The LIGO design case revealed a number of concepts that are important elements of creative visualization interactions. Tacit and explicit knowledge (Collins, 2007; Miller, 2008; Polanyi, 1966) turned out to be useful framing concepts for understanding the differences between information that different individuals bring to a creative visualization exercise themselves vs. the information that may already be codified in the data to be visualized. Underlying data from the LIGO computational pipeline explicitly contained important information to be visualized, but the scientists who use this pipeline and originally designed various aspects of it also tacitly held knowledge that was critically necessary in order to meaningfully visualize the entire computational process. Only in combination could tacit and explicit information actually be used to produce meaningful visual artifacts.

Intermediary visual artifacts also emerged as a key aspect of the creative visualization interaction. LIGO scientists, IR researchers, and IV researchers all collaborated to produce a variety of rough sketches, incomplete diagrams, and temporary chalkboard drawings of the LIGO pipeline. Though inaccurate in and of themselves, these images were critical for orienting participants and enabling further discussions of the visualization activity at hand, a process known as image-enabled discourse (Snyder, 2009a, 2009b, 2012a, 2012b, 2012c).

This design case details how, over the course of a participatory research and design process, creative visualization interactions were critical to the development of visualization tools for LIGO. Indeed, it demonstrates how under some circumstances these kinds of human-centered approaches to visualization are not merely desirable, but actually critical to the ultimate achievement of the visualization task.

Problem Overview

The LIGO Scientific Collaboration

The Laser Interferometer Gravitational Wave Observatory (LIGO) is an international scientific collaboration to detect gravitational waves that can be generated in a variety of ways (e.g. 'inspiral' waves produced when two closely orbiting stars or black holes collide and merge with each other, or 'stochastic' waves, which are the detectable effects of the cosmic gravitational wave background) (Barrish & Weiss, 1999; Owen et al., 2011). To date, no gravitational wave has yet been directly detected.

Laser sensor systems, called Michelson interferometers, are used to detect gravitational waves. The LIGO interferometers consist of two lasers in a vacuum, arranged perpendicular to each other in an 'L' shape and pointed down very long (4km) 'arms' at corresponding photodetectors. Gravitational waves can have a measurable physical effect on these laser instruments, and data from several of these instruments, located at various locations in the United States and around the world, is typically combined for analysis (Abbott et al., 2004; Barrish & Weiss, 1999). Because of the extreme sensitivity of the laser interferometer instrumentation, the data produced is very noisy. For example, a member of the LIGO collaboration indicated that a train passing several miles from a detector would be enough to render instrument data from that time unusable; interference effects combined with gravitational wave signals that are faint to begin with produce a low signal/noise ratio (Barrish & Weiss, 1999).
To address the signal/noise issue, LIGO scientists use a complex computational pipeline running on multi-CPU/GPU server clusters (2011) to filter the large amounts of raw data available from the interferometer instruments. These computer processes include steps to ‘veto’ bad data, look for ‘triggers’ (signals that may be worth examining more closely), match triggers to coincident data from other detectors, and evaluate whether a trigger actually matches the signature of a gravitational wave event. The process of filtering signal from noise within LIGO data is computationally expensive because of the quantity of data and the many steps needed to process it. LIGO scientists have an interest in avoiding duplicated computational runs by producing information tools that 1.) inform other collaboration members of what interferometer signals have previously been processed and how, and 2.) provide ways for collaboration members to select and incorporate portions of previously completed processing runs into newly established processing runs.

Information Retrieval Task

An information retrieval (IR) tool to search previous computational processing activities was proposed. Members of the [Institution Omitted for Review] Gravitational Wave Group (a member group of the LIGO scientific collaboration looking specifically for ‘inspiral’ gravitational wave events) therefore began working closely with IR researchers to produce such a tool.

Visualizations of the computational pipeline itself soon became an intended component of the proposed IR tool. Such visualizations were expected to be useful for several reasons. Scientists using the IR system could use the returned visualizations of previous computational runs to help make decisions about the value of those runs for their own work (use case 1). Furthermore, such visualizations could eventually be developed into interfaces for LIGO scientists to easily select portions of previous runs as a basis for future work (use case 2).

It is understood that knowledge construction and communication can be enabled by visualization tools (Fischer et al., 2002), as well as sketches or drawings produced during conversation, a phenomenon sometimes called image-enabled discourse (Snyder, 2009a, 2009b). Even rough, incomplete, or inaccurate diagrams of LIGO’s complex computational processes would be expected to enable more detailed discussion, and thus the generation of progressively more detailed and accurate visualizations and a more useful IR system overall. Though not considered to be a final deliverable for this project, intermediary visualization artifacts became an important incidental use case (use case 3).

Visualization Task

Use cases 1 and 2 were dependent on successfully developing visualizations for use case 3. That is, before dynamically generated visualizations depicting specific computational runs could be produced for the IR system, static, hand-crafted visualizations of the pipeline in its general form would be needed to improve communication and planning (e.g. preparation for the tool to be built). The majority of this design case addresses visualization artifacts produced for use case 3.

Challenges

Creating visualizations for case 3 was a non-trivial task. The LIGO computational pipeline is highly complex. It is documented using a wiki system that relies on Gravitational Wave Group scientists’ own initiative to contribute and has very little overarching structure or organizational control. The wiki website describes pipeline functionality, but not a systematic or complete way. Most experienced Gravitational Wave Group scientists become familiar with the pipeline through their use of it over time, and the executable files of the system itself become the main form of documentation. This is problematic for new inductees to the project, including new researchers and graduate students. It was similarly problematic for the IR and IV researchers involved in planning and developing IR and visualization tools, who had no knowledge of how the pipeline worked. For the Gravitational Wave Group scientists themselves, reliance on executable files and their own knowledge can break down when the need arises to undertake infrequently performed tasks.
IR researchers, having no familiarity with LIGO or the computational pipeline, began by automatically producing a diagram of the pipeline, drawing upon the many thousands of intermediary files that the pipeline executes during a run. Executables within the pipeline have multiple input and output files, and these could be visually reproduced, along with their connections to one another.

*Figure 1. The first LIGO pipeline visualization (< 1% of total visualization shown).*

In most respects, this first visualization attempt was a failure. It was massive in scope, and overwhelmingly complex. As an automatically generated diagram based on simple input/output links between executables, it was also an exercise in pure representation that excluded any kind of human interaction with the data or the visual output. The diagram was static and printed on paper. It had no affordances for representational interactions. Users could not easily change or manipulate links between nodes, seeking patterns or meaningful connections. Furthermore, though this visualization did support cognitive interaction, its complexity and scale acted as more of a barrier rather than an aid to understanding. This is to say that the diagram’s massive and seemingly patternless network of links and nodes, as well as the fact that no human knowledge of the pipeline and its workings had helped to shape it, conspired to generate more confusion than clarity. It was clear from discussions with Gravitational Wave Group scientists that some underlying organizational structure had been excluded from the visualization, resulting in its misleading scale and complexity, but it was unclear as to what this structure should look like or how it could be visualized.

**Method**

Having failed with an automatically generated visualization, IR and IV researchers directed their attention toward human-centered visualization – that is, creative visualization interactions. By seeking rich knowledge of LIGO practices and activities and creating different visual artifacts to foster discussion, IR and IV researchers hoped to unlock the underlying structure of the LIGO pipeline. Participant observation, a qualitative, action-oriented method of research, would be used. DeWalt and DeWalt (DeWalt & DeWalt, 2002) describe this method as follows: "Participant observation is a method in which a researcher takes part in the daily activities, interactions, and events of a group of people as one of the means of learning the explicit and tacit aspects of their life routines and their culture." This method, "draws on the insights gained through the use of participant observation for gaining greater understanding of phenomena from the point of view of participants," (DeWalt & DeWalt, 2002)
Active researcher participation of this kind is sometimes frowned upon because of the potential to introduce researcher bias (Yin, 2009). However, active involvement can also be beneficial and preferable in some cases. According to DeWalt and DeWalt (DeWalt & DeWalt, 2002), "The participating observer seeks out opportunities to spend time with and carry out activities with members of communities in which he or she is working. Because enculturation takes place at the same time (it is hard to avoid), we believe that a tacit understanding of the experience is also being developed. It is an understanding that is not easily articulated or recorded, but that can be mobilized in subsequent analysis."

A series of ten meetings were held over the course of three months with Gravitational Wave Group members and IR/IV researchers. Gravitational Wave Group members attended six of these meetings, while IR and IV researchers were present at all ten. Meetings were typically between 2 and 4 hours in length.

Gravitational Wave Group participants included two scientists (a gravitational wave physicist and a high performance computing specialist), each of whom had high levels of experience in the use and design of the computational pipeline. In addition, three graduate students in physics participated in one meeting; their experience ranged from high levels of proficiency (a senior student who was close to defending his dissertation) to intermediate (students who had been in the physics program for a few years). IR group participants included the director of [Institution Omitted for Review] Center for Natural Language Processing (CNLP), as well as a technical specialist and programmer who would ultimately be directly responsible for developing the IR system. The author of this design case, an IV researcher, was also considered to be a member of the IR group because of the close relationship between the visualization task and other IR-specific tasks.

In accordance with participant observation and action-oriented research norms, IR/IV researchers did not assume a purely observational role, but rather were partners in the project, fully engaged contributors to the various topics under discussion. They guided conversations to points where knowledge held by the Gravitational Wave Group scientists could be extracted and codified, and sought to understand LIGO activities through a collaborative design experience. For example, IR/IV researchers participated in instantiating a LIGO pipeline "run," choosing run parameters and learning from the physicists who normally undertake such work. By participating in normal pipeline operations themselves, IR/IV researchers had a frame of reference for how the pipeline system worked and what kinds of information and decisions went into a successful run.

Detailed notes in the ethnographic research tradition (Emerson et al., 1995) were taken at all meetings. In addition, photographic records were made of important visual artifacts resulting from meeting discussions. These included chalkboard and whiteboard notes, sketches, diagrams, and hand-drawn notations made on previously produced visual material. Gravitational Wave Group and IR/IV researchers also exchanged regular emails that expanded upon and clarified information discussed during meetings. These emails became an additional source of data.

Findings

The LIGO Pipeline Overview

Even when the decision was made to abandon the preliminary, automatically-generated visualization, progress seemed not to have improved. One reason was the use of specialized terminology by Gravitational Wave Group members. LIGO and its associated organization, instrumentation, scientific goals, and the pipeline itself are complex, so it took time for IR and IV researchers to become comfortable with a glossary of highly specialized terms: "DAX," "DAG," "Condor," "Pegasus," "veto," "cat," "event," "frame," "workflow," etc. More challenging, it soon became apparent that members of the Gravitational Wave Group defined these terms differently. No standardized glossary of terms existed, so each scientist used a mixture of commonly-understood and personalized terminology. This led to many disagreements and discussions about specific meanings for a host of different terms.

Members with high levels of experience held relatively accurate views of the pipeline, while newer group members had less accurate and sometimes underdeveloped views of the pipeline. One PhD student at an intermediate experience level described the pipeline as a "black box" and described previous visualization efforts that emphasized only the computational aspects of the pipeline as "next to useless." Yet thoroughly understanding the pipeline and its various components and relationships was important to this student: "Knowing what this [the pipeline] is doing... that is the biggest thing." An
advanced doctoral student further indicated that, "Intermediate data products are indecipherable until you know the whole system." Similarly, IR and IV researchers needed terminology to be semantically defined in detail, but also depicted visually in relationship to other terms and components of the pipeline. No such visualization existed, so after four meetings, IR and IV researchers attempted to make one on their own, producing a whiteboard diagram over a two-hour discussion session.

Despite the risks of attempting a visualization without Gravitational Wave Group members in attendance, this intermediary visualization artifact was a moment of clarity for the IR and IV researchers who had produced it. Since it had been produced by members of the collaboration for themselves (rather than automatically generated or provided by outsiders), it was simpler and clearer than previous automated attempts, uncluttered by masses of irrelevant data. It was also a representation of what IR and IV researchers thought they knew, rather than a representation of whatever was possible to produce from pipeline data alone. This distinction turned out to be invaluable.

Nonetheless, the visualization highlighted key doubts held by the IR and IV researchers, as indicated by quotes from the meeting where it was produced: "I think that is what they were saying," and, "Does this [diagram] seem to resemble what we have been talking about?" Some nodes of the pipeline structure were highly uncertain. For example the processes "Condor" and "Pipedown" were known to be inaccurately represented, but were depicted as they were because of the need to create some foundational visual artifact for discussion.

An identical (albeit tidier) version of this diagram was prepared digitally and printed for discussion in a follow-up meeting with Gravitational Wave Group members. This "discussion version" of the diagram had an immediate effect. Because the diagram was not a depiction of truth or data, but merely an attempt to make sense of many prior discussions, collaborators felt free to mark and correct it at will. IR and IV researchers knew the diagram was inaccurate, but hoped that Gravitational Wave Group members could identify those specific inaccuracies and correct them. This turned out to be the case, as the collaborators held a visually-enabled discussion, sometimes referred to as image-enabled discourse (Snyder, 2009a, 2009b, 2012a, 2012b, 2012c). Some aspects of the diagram were certified as accurate, but many others were revised:
Figure 3. LIGO pipeline overview (discussion version).

The overview visualization and subsequent discussion established that the processes called "Pegasus," "DAGMan," and "Condor," interacted in the pipeline to parse through an XML formatted hierarchy of interrelated computational jobs (a "DAX"), establish job dependencies, and feed those jobs to processing resources. This highly complex series of events was made more so by the addition of antecedent processes such as "iHope" that form the DAX based either on user input or the "cache" where users can substitute parts of previous runs into the pipeline data to avoid unnecessary duplication of computational effort.

A major flow of the automatically produced diagram was that it could not show variations that occur between different pipeline runs. Each run produces data only about itself, and the automated visualization was produced out of data from just one run. In contrast, the discussion diagram was a holistic view produced from human understanding of the pipeline. It showed the activities surrounding the pipeline, including human activities that are not computational at all. It also showed various points where modifications or data insertions might be made by Gravitational Wave Group members. Though it took many hours of hard work, discussion, and correction to achieve, the visualization produced through creative visualization interactions among several collaborators was a more useful tool by far than the automatically generated visualization. A senior member of the Gravitational Wave Group described the overview visualization as, "probably the clearest picture we've ever had of this process."
The pipeline overview was a necessary, but not sufficient, first step toward producing visualization-enabled IR tools that could accommodate the typical functionalities of overview, filter, and detail (Shneiderman, 1996). Since all members of the collaboration now had a reasonable understanding of the basic pipeline structure, conversation turned to the details of the “DAGman” and “Condor” processes. In the overview diagrams, this portion of the pipeline was initially represented by a series of nodes and arrows depicting the flow of data through a series of interdependent computational jobs. It was not yet clear how this specific structure was organized.

**DAGman and Condor**

The process for unpacking “DAGman” and “Condor” was similar to that for creating the overview visualization. A series of discussions led to informal visualization attempts (usually during meetings), followed by collaborative corrections and a final, refined diagram. Perhaps the most critical piece of information discerned in discussions about “DAGman” and “Condor” was the role of time in the interferometer data itself.

The LIGO interferometers, when turned on, provide a steady stream of data, similar in some respects to an audio microphone. Analyzing this data requires that scientists define a past timeframe to analyze, be it just a few seconds, several minutes, or even hours at a time. The interferometer data is subdivided into frames, consisting of 16 seconds each, and these frames constitute the minimum actionable unit of data that is addressed by jobs within the pipeline. Members of the Gravitational Wave Group eventually codified this via a chalkboard visualization showing how the interferometer data became partitioned into frames. This visualization was produced over the course of a two-hour collaborative discussion:

---

**Figure 4.** The final LIGO pipeline overview visualization.
Figure 5. Interferometer data represented as 16 second frames.

This drawing led to the key finding for this design problem: that the LIGO pipeline analyzes many frames of data simultaneously, and these frames are all subjected to the same relatively simple computational process. A second chalkboard diagram showed how each frame of data would be submitted to a computational node (labeled as “A”). From there, output data would be submitted to two new nodes (“B” and “C”), and their output then combined by a fourth node (“D”).
Figure 6. DAGman and Condor in chalk, used to enable discussion.

Again, a creative interaction was necessary to make this organizing structure fully clear. One of the Gravitational Wave Group members described the ‘fatal flaw’ of the initial, automatically-generated visualization: “the graphing code can't figure out that there are a bunch of nodes that are really the same job repeated n times with different GPS start and end times.” In this instance, human input to the visualization activity was necessary to reveal both the problem and its solution.

Interferometer frames were one important dimension of visualizing the pipeline. A second key dimension was the number of interferometers that are included in a run. There are several interferometer instruments located around the world, including instruments that are no longer operational but have archived data sets that are still useful. LIGO scientists can choose to include data from one or several of these instruments in a given run (typical runs use more than one data source to more easily filter signal from noise). The pipeline subjects each frame of data from each interferometer to the same series of computational transformations, so the number of interferometers examined also affects how many of each type of job will occur.

The third key dimension for visualization was known from the start: processing time. Frames of data progress through the LIGO pipeline in a relatively linear fashion, with the output of one job acting as an input to others. The time it takes for frames of interferometer data to progress through the pipeline is dependent on the power of the computational hardware that the pipeline is running on. Jobs that occur later in the pipeline are generally not able to be started until previous jobs have completed.

With knowledge of these three dimensions for visualization - frames, number of instruments, and processing time - IV researchers manually produced a new 3D diagram showing an abstracted view of the pipeline structure.
Figure 7. The pipeline in three dimensions: frame, instrument, and processing time.

This new diagram captured the essence of the pipeline structure, though it encompassed only part of the full process and lacked accuracy in several respects. Over the course of the remaining discussions it became apparent that visualizing the frame dimension of the pipeline (e.g. its 3D “depth”), while useful as a conceptual aid for those unfamiliar to the pipeline, was unimportant to project scientists in their real-world activities. A finished visualization tool for LIGO could collapse this dimension, the LIGO scientists being better served by numerical representations indicating how many frames had been processed rather than visualized depth.

The final visualization produced for the LIGO design case was a version generated automatically from real data, this time incorporating the knowledge gathered over the course of many round of creative visualization activity: the frame, instrument, and time-to-process dimensions, the pipeline structure, and the final use scenarios for the visualization itself.
Figure 8. The computational pipeline.
Discussion

What do groups of people gain by visualizing information for themselves? The LIGO design case reveals a number of interesting possibilities that are likely to be generalizable to other visualization problems.

Cooperative Sensemaking

Many prominent IV researchers (e.g. Chen, 2003, 2006; Spence, 2001, 2007; Ware, 2004) emphasize the benefits of visualizations for human cognition. The basis for most cognitive theories of IV is simple: the human mind is not very good at holding many objects in short term memory all at once. IV tools are an aid to human memory, visually storing objects for us and arranging them in a fashion that will improve our ability to understand and draw conclusions about them (S. K. Card et al., 1999). Spence (2001) describes IV as the formation of a mental model through cognitive interaction with a visual display of data. Mental models help users to form a, “better understanding of the artifact, scheme, or situation to which the data refers, and to be able to interpret that model in some useful way, perhaps to make a decision,” (Spence, 2001). Ware (2004) emphasizes the perceptual aspects of cognition, describing many of the ways human beings see and interpret different kinds of visual data. The notion of image-enabled discourse (Snyder, 2009a, 2009b) is related to this, conceptualizing, as it does, visualization artifacts as communication tools (Fischer et al., 2002).

These benefits certainly applied to the visualizations produced by IR and IV researchers. These individuals came to the collaboration with virtually no knowledge of LIGO, the computational pipeline, subprocesses within the pipeline, or any other aspect of the gravitational wave research being undertaken. Gravitational Wave Group members and non-physicist scholars found themselves talking at cross-purposes, mixing vocabulary, and misunderstand fundamental terminology and concepts as they strove to understand core elements of the problem at hand. Four meetings passed with minimal progress made; IR and IV researchers couldn’t see beyond the massively complex and disorganized early visualization of the pipeline, while Gravitational Wave Group members seemed to be suggesting that in reality, the pipeline was far more structured and simple than the diagram implied.

By producing a series of visualizations for themselves, physicists, IR, and IV researchers managed to bring sense to something so complex as to seem senseless. Intermediary and even inaccurate visualization artifacts served an important purpose, anchoring discussion and directing it toward the improvement of the diagram. This is a key benefit of intermediary visualization artifacts and of cooperatively creating visualizations: the visualization and the act of creating it are mechanisms to explore the knowns and unknowns of the underlying information. The act of creation requires the different individuals involved to cognitively interact with data, filtering and manipulating it, contributing to it, exploring it, and ultimately representing it in a visual form that makes sense to everybody involved.

Codifying Tacit Knowledge

In the LIGO design case, as with many visualization activities, data manifested in two forms. Some data was concretely represented and accessible, such as pipeline input and output files. Other data was known, often in partial form, only by certain individuals. This data was not concretely represented or accessible to others, even in existing project documentation. It proved helpful to conceptualize these two types of data through the constructs of explicit and tacit knowledge (Collins, 2007; Miller, 2008; Nonaka, 2007; Nonaka & Takeuchi, 1995; Polanyi, 1966).

Explicit knowledge is codified knowledge, information that has been recorded or can be observed and explained in some concrete fashion (for example, mathematical formulae or a written record of a person’s observable daily routine). Tacit knowledge is personal and specialized. It is un-codified knowledge that resides only within the individual and may be difficult to translate into explicit form. Because of its personal nature, tacit forms of knowledge can only be accessed by human beings (Collins, 2007; Miller, 2008; Nonaka, 2007; Nonaka & Takeuchi, 1995; Polanyi, 1966). For example, Collins (Collins, 2007) describes riding a bicycle as an expression of tacit over explicit knowledge: it is possible to read and understand codified explanations of the physics and mechanics of riding a bicycle, but this is not the same as knowing how to ride a bicycle. The physics and mechanics are explicit knowledge, while the understanding of how to ride is tacit knowledge.
In visualizing the pipeline, matters of explicit vs. tacit knowledge were highly important. Computational artifacts (input, output, and executable files) were explicitly documented by the pipeline as it ran. These artifacts contained a great deal of information about the pipeline itself: file names reflected important details about relationships to or dependencies upon other files, as well as information about which 'jobs' (executable code) the various input and output files were associated with. However, early attempts to visualize the pipeline using only explicit, codified information showed that deeper knowledge of the pipeline would be required to produce anything meaningful. IR and IV researchers knew from early conversations with Gravitational Wave group physicists that the pipeline had an underlying structure that was somehow not being reflected through use of explicit, data-oriented knowledge alone. Visualizing this underlying structure required that IR and IV researchers gain access to the tacit knowledge of it held by the Gravitational Wave Group physicists. This tacit knowledge included details of the pipeline that were not, but could have been, made explicit. It also included knowledge of how to execute a computational run and how to read, interpret, and use results in the overarching context of the LIGO organization. A model for using visualizations to connect tacitly held knowledge to explicitly held knowledge has been described by (Jeong et al., 2008), though this model assumes a conversion of explicit knowledge to tacit knowledge. In the LIGO visualization task, this flow was reversed; the design process extracted tacitly held knowledge from a variety of collaborators and structurally applied it to explicit knowledge sources through visualization.

Intermediary artifacts (the pipeline overview diagrams; the chalk sketches of "DAGman" and "Condor") allowed all members of the collaboration to begin explicitly codifying their interpretation of information under discussion. Even when these visual interpretations were incorrect, the fact that they were codified made it far easier for others with different – possibly more accurate – interpretations to weigh in and discuss corrections. This visually-enabled process collaboratively improved all participants’ tacitly held knowledge and simultaneously resulted in an accurate and explicitly codified visual representation of that knowledge.

**Beyond Data, Technology, and Technique**

In the current IV literature, much has been made of data, technology, and technique as determiners of a visualization’s success or failure. Topics of interest include how different kinds of data or the scale of a data set (e.g. Aigner et al., 2007; Fekete & Plaisant, 2002; Keim, 2001; Shneiderman, 1996) require different approaches to visualization. 3D technologies (Robertson et al., 1993; Robertson et al., 1991), graphics processing hardware (e.g. Rhyne, 2000), and mobile platforms (e.g. Chittaro, 2006) are also of interest. Finally, many visualization scholars have explored different approaches to visualization, including "Fisheye Views" (1986), virtual worlds (1990), 3D rooms (1991), “Cone Trees” (Robertson et al., 1991), and many more.

The LIGO design case reveals how this tripartite scholarly emphasis on data, technology, and technique frequently overlooks an important fourth dimension: human-centeredness and the very act of creating visualizations. In this design case, neither data nor technology proved very helpful by themselves in achieving a meaningful and useful visualization; indeed, in some respects these were an obstacle to progress, as massive amounts of disorganized data and automated visualization techniques added to the confusion of the collaborators. In the same way, specific visualization techniques were not of particular interest. A variety of techniques and media were used: whiteboard, chalkboard, paper + pen, 3D, and 2D. The selection of these was made based on convenience and availability and was not, per se, the driving force behind the visualizations that were produced.

On the other hand, the direct involvement of many different collaborators in the creation of visualizations for themselves was a critical and important element. Only by casting aside automated shortcuts to deeply immerse themselves in the Gravitational Wave Group’s activities, difficulties, and objectives, were IR and IV researchers able to begin composing visualizations with any real utility. Ultimately, it was enhanced human understanding of the pipeline and the objectives of the Gravitational Wave Group itself that led IR and IV researchers back to automatic visualization, this time with highly successful results.
Conclusion

Creative interactions with visualizations can have many benefits and advantages. In this design case, collaborators worked together to build several visualizations of a complex computational pipeline. Their process helped to transform tacitly held knowledge into actionable forms of explicit, codified knowledge. Without this transformation process, generating meaningful visualizations of the LIGO pipeline would not have been possible. As such, the visualizations generated over the course of this study served as both completed interim products of research and as tools to foster image-enabled discourse for future tool planning and development.

References


Scientific Referential Metadata Creation with Information Retrieval and Labeled Topic Modeling

Han Jia  
School of Library and Information Science  
Indiana University  
hanjia@indiana.edu

Xiaozhong Liu  
School of Library and Information Science  
Indiana University  
liu237@indiana.edu

Abstract

The goal of this research is to propose an innovative method of creating scientific referential metadata for a cyberinfrastructure-enabled learning environment to enhance learning experiences and to help students and scholars obtain better understanding of scientific publications. By using information retrieval, topic modeling, and meta-search approaches, different types of resources, such as related Wikipedia Pages, Datasets, Source Code, Video Lectures, Presentation Slides, and (online) Tutorials, for an assortment of publications and scientific (labeled) topics will be automatically retrieved, associated, and ranked. In order to test our method of automatic cyberlearning referential metadata generation, we designed a user experiment for the quality of the metadata for each scientific keyword and publication and resource ranking algorithms. Evaluation results based on MAP, MRR, and NDCG show that the cyberlearning referential metadata retrieved via meta-search and statistical relevance ranking can effectively help students better understand the essence of scientific keywords and publications.

Keywords: metadata generation, information retrieval, referential metadata, cyberlearning resource, user, scientific publication, labeled topic modeling

Introduction

In the past decades, rapid access to digital publications accelerated and facilitated study and research; however, several challenges should be addressed. First of all, the sheer volume of scholarly publications available online makes it impossible for a researcher to absorb all the new information available. Hence, researchers and students need to find innovative ways of quickly and effectively learning and understanding new scientific topics and publications. But existing tools are only limited to descriptive information about publications of a specific topic. Second, understanding the content of scientific publications remains daunting. For instance, in a recent survey we conducted with students in a class on “information retrieval theory and practice”, the complex models, algorithms, formulas, and methodologies in the publications were often found too difficult to understand due to their limited backgrounds in computer science, statistics, and mathematics. Third, some recent exciting developments have illustrated the possibility of utilizing multimedia content—i.e. videos and images—to facilitate students and scholars to understand scientific content. However, the cost of generating sophisticated cyberlearning resources for large-scale scientific topics or publications makes the approach prohibitive.

Metadata have traditionally centered on descriptive representation through title, author, publisher, subject keywords and other attributes of scholarly output. Descriptive metadata, however, have become increasingly inadequate as the complexity and volume of scholarly output grows. Innovative mechanisms have been developed to address these new challenges, for examples, Liu, et al. (2011) and Liu (2012) proposed referential metadata, which refers to information about any sources implicitly or explicitly cited in a publication or about artifacts associated with the publication. Referential metadata provides a context, in which the publication was created, the co-authors who collaborated, and the information and data that were used. They also link to the artifacts that the publication may have generated: a presentation video or slides, images, datasets, tutorial, source codes, and even question-answering documents. The referential
metadata may not all exist in the publication as references and some may be scattered across the researcher’s project or personal website or social media, i.e. TED, SlideShare, Sourceforge, and Wikipedia, making it more difficult to obtain them by conventional metadata generation methods. However, referential metadata can effectively help readers to better understand the essence of the publication. As detailed in the figure below (Figure 1), in order to help students and scholars understand such complex publications, it is critical to offer them resources that may enhance their learning process, such as user-friendly tutorials and video lectures.

![Figure 1. Depiction of cyberlearning resources related to a scholarly publication.](image)

For this research, instead of generating resources (referential metadata) for each topic or publication, we assumed that cyberlearning resources were available on the Internet and tried to automatically retrieve, associate, and rank resources based on their “importance” for each scientific topic and publication. In other words, we conceptualized the problem of creating cyberlearning referential metadata as an information retrieval problem that amenable to automation using meta-search and retrieval algorithms. The retrieved resources were ranked based on the content relevance (language model), topic relevance (labeled LDA inference), and a combination of the two factors (topic probability as language model prior).

In order to test our method of automatic cyberlearning referential metadata generation, we designed an experiment in which a group of graduate students learn and understand the essence of randomly sampled research topics and publications in the information retrieval domain through automatically generated cyberlearning resources. Evaluation results show that automatically generated cyberlearning resources via retrieval and meta-search can effectively help students to understand the essence of scientific topics and publications. In addition, we utilized student feedback to validate the
ranking algorithms designed to prioritize informative resources for the target scientific topics and publications.

In the remainder of this paper we (1) review relevant literature and methodologies, (2) introduce our referential metadata generation methodology, (3) describe our experiment in the information retrieval domain with a group of graduate students with respect to both sampled topics (keywords) and publications, (4) evaluate the cyberlearning resources referential metadata creation and resource ranking algorithms, and (5) discuss the contributions and limitations of our work.

Previous Research

Referential metadata (Liu, Chen & Qin, 2011) in the broadest sense include not only citations but also data about other types of scholarly output that is based on or related to the same publication. It is common today that before a paper is published in a journal, the authors of the paper may have presented it as a conference poster and/or a conference paper which produces presentation files or videos, or have made datasets, source code, or related materials available on the project website. These precursors and artifacts of a publication establish a context as well as provenance for readers to understand, evaluate, and interpret the research reported in the publication. While referential metadata are valuable for information retrieval and use, they are not usually included in publication metadata records.

Cyberlearning resources, as a specific kind of referential metadata, are highly important for e-learning environment. In this research we focused on creating referential metadata for some specific scientific topics and publication. In this section we review previous efforts on scientific metadata creation for cyberlearning resources.

Cyberlearning Resources and E-Learning Challenges

The proliferation of cyberinfrastructure and resources calls for more powerful and effective metadata representation methodologies to address information discovery and e-learning challenges. Referential cyberlearning metadata, as an emerging effort devoted to providing scientific topic- and publication-rich web context, is necessitated by the exponential growth of online open resources. The commitment of researchers in this field to education and e-learning should not be ignored.

Nevertheless, online open resources typically lack clear quality assurances. This is now recognized as a major concern with online open resources (D’Antoni, 2009). Not until fairly recently have researchers used multimedia Web resources, such as videos, audios and images, as an effective means of supporting student learning. For example, DeLeng, Dolmans, & van de Wiel (DeLeng, Dolmans, & Wiel, 2007) examined students’ views on the added value for problem-based learning (PBL) of using video resources in contrast with exclusively text-based approaches during the pre-clinical phase of undergraduate medical education. In the experiment with undergraduate students, they found that videos were generally perceived as a valuable stimulus for group discussions in PBL. Similarly, Maniar et al. (2008) examined the possibility of using mobile phones for video-based learning, a.k.a. m-learning.

Persson, Fyrenius, & Bergdahl (2010) used multimedia resources to enhance problem-based learning across the entire curriculum, making education more realistic and thereby more motivating and stimulating for students. Furthermore, Agazio and Buckley (2009) used YouTube for nursing education. In their research, YouTube was used to illustrate theoretical content, involve students, and inspire innovative teaching methods. Videos were presented on YouTube to stimulate student discussion, share information, and create a learning community. Duffy (2008) similarly, investigated the possibility of using YouTube, Podcasting, Blogs, Wikis and RSS to create a ubiquitous user-centric, user-content generated and user guided learning experience.

Typically, however, it is extremely costly to generate sophisticated cyberlearning resources for specific scientific topics and publications. Consequently, there is a need for a more effective and efficient set of tools and methods to create, associate, and manage online cyberlearning resources. To the best of our knowledge, there are few studies focusing on the task of bridging the gap between existing resources and scientific topics and publications.
Scientific Metadata for E-Learning

Metadata has traditionally been used for finding, identifying, selecting, and obtaining information objects. In its short history, however, metadata research has split into two camps with different perspectives and paradigms: 1. The description paradigm found in library and information science; 2. The processability and executability paradigm rooted in computer science (Zeng & Qin, 2008). Research on metadata representation and generation over the last few decades has drawn techniques and methods from a wide variety of research fields, including natural language processing and machine learning.

Obviously, for most existing library or document repository systems, professional metadata creators or domain experts (e.g., catalogers and indexers) are the ideal candidates (Milstead & Feldman, 1999) to create metadata, as they are familiar with the systems and terminologies. However, this approach is costly and may be limited in availability. It is hard to apply this approach to large amounts of data across different domains. Other researchers, for instance, Greenberg et al. (Greenberg, Pattueli, & Parsia, 2001), have found that authors can sometimes provide higher quality metadata for web resources. This approach is adopted by most digital libraries. Nevertheless, many authors are only willing to provide relatively simple descriptive metadata. In the medical domain, researchers have found that explicit structural abstracts are not entirely reliable (Demner-Fushman & Lin, 2007).

For these reasons, user- or author-generated and professional- or expert-generated metadata can hardly cope with the need for complex referential metadata generation at a large scale across different domains. Accordingly, the automatic approach, an economical and effective alternative, has become popular over the past few years. A wide variety of techniques have been used to process digital texts to generate metadata records. Some existing meta-search engines, such like NECI (Lawrence & Giles, 1998), SavvySearch (Dreilignner & Howe, 1996), and MetaCrawler (Selberg & Etzioni, 1995), have proved meta-search a successful technology for enhancing the user search experience. Unfortunately, generating metadata for cyberlearning resources for research topics and publications that were neither directly created by authors or publishers nor explicitly referenced in the content of the papers is a very demanding task.

Tremendous efforts have been made to improve the quality and efficiency of search engines over the years. However, relying on only one of them is insufficient. Meta-search is an information retrieval method that sends queries to multiple search engines or digital libraries simultaneously and aggregates the results into a single list. The basic assumptions are that a single search engine can only index a small portion of the web, and that aggregated retrieved results are likely to be more comprehensive. Some existing meta-search engines, such like NECI (Lawrence & Giles, 1998), SavvySearch (Dreilignner & Howe, 1996), and MetaCrawler (Selberg & Etzioni, 1995), have proved meta-search a successful technology for enhancing the user search experience.

Cyberlearning Referential Metadata Creation

As domain knowledge in most disciplines expands at a frenetic pace, disconnected research artifacts and other resources need to be connected in innovative ways, as through more effective use of metadata. In this research referential metadata were generated for six different types of cyberlearning resources—Wikipedia pages, video lectures, (presentation) slides, datasets, (online) tutorials, and source code—to help students and scholars better understand the scientific topic and publication.

In this section, we will explain our approach through three individual questions step by step: 1) How to collect different types of resources for a certain topic; 2) How to perform topic modeling on publications; 3) After gathering resources and differentiating topics, how to associate resources with each publication by using topics as intermediary and rank these resources.

Collecting Resources for Topics

In this research, we treated author-assigned keywords as a representation of each scientific research topic in a given domain of study, namely, information retrieval. In order to generate referential metadata for each topic effectively, information retrieval and meta-search approaches were applied, in which a Boolean query was sent to one or more search engines for each scientific keyword and
cyberlearning resource type. The detailed query for each resource type is listed in Table 1. The first column is the cyberlearning resource type. For each type the target query was sent to one or more search engines. For example, in order to get tutorial resources for a keyword, the query “[Keyword] AND Tutorial NOT Video” (column 3) was sent to Google (column 2), where [Keyword] was replaced with the target keyword content. Similarly, for video resources, the query “[Keyword]” was sent to YouTube1, TED2, and Videolecture3. For slides and source code, different queries were sent to different search engines. For example, for slides the queries “[Keyword] AND (filetype:ppt OR slides)” and “[keyword]” were sent to Google and Slideshare4, respectively. For performance reasons, we indexed the Wikipedia page dump locally.

Table 1
Meta-search query for different kind of resources

<table>
<thead>
<tr>
<th>Resource type</th>
<th>Search engines</th>
<th>Query (Keyword task)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia page</td>
<td>Wikipedia dump (local database)</td>
<td>[Keyword]</td>
</tr>
<tr>
<td>Tutorial</td>
<td>Google</td>
<td>[Keyword] AND Tutorial NOT Video</td>
</tr>
<tr>
<td>Slides</td>
<td>Google, Slideshare</td>
<td>Google: [Keyword] AND (filetype:ppt OR slides) Slideshare: [Keyword]</td>
</tr>
<tr>
<td>Video</td>
<td>YouTube, TED, Videolecture</td>
<td>[Keyword]</td>
</tr>
<tr>
<td>Dataset</td>
<td>Google</td>
<td>[Keyword] AND Dataset</td>
</tr>
<tr>
<td>Source code</td>
<td>Google, SourceForge</td>
<td>Google: [Keyword] AND (Source Code OR Toolkit OR Java OR C++ OR Python) SourceForge: [Keyword]</td>
</tr>
</tbody>
</table>

Based on the Table 1, a list of queries was sent to different search engines to retrieve candidate resources. In our experiment, we used the top 15 retrieved results from each search engine to aggregate the final result collection for each resource category. In most cases, the result collection was a combination of informative resources and noisy results. Experience in information retrieval reveals that since different search engines return very diverse and sparse results for the same query or for a similar one, irrelevant data may pollute the search results and mislead users. For instance, as the following diagram shows, for the topic labeled “Question Answering,” if we use the topic label as our query, two Wikipedia pages get high content relevance scores (i.e. BM25 or TF-IDF). However, users are only interested in the first one, “Open domain question answering,” for this scientific topic, while the second one, “Question and Answer (album),” should be removed or ranked lower as a noisy resource. Current information retrieval ranking methods based on bag-of-words, like language model and BM25, can hardly detect the topic level match.

To address this problem, we used the topic modeling algorithm to generate the word probability distribution for each scientific topic (keyword). For each scientific keyword, keyi, a topic-word distribution \( P(\text{word}_k|z_{\text{key}_i}) \) needs to be trained from scientific literature. Then, we can enhance the ranking algorithm to prioritize those informative resources for the target scientific keywords by using resource topic prior probability. In more detail, if we use \( P(\text{resource}|\text{key}_i) \) to rank each candidate resource:

\[
P(\text{resource}|\text{key}_i) = \frac{P(\text{key}_i|\text{resource}) \cdot P_{\text{prior}}(\text{resource})}{P(\text{key}_i)}
\]

---

1 [www.youtube.com](http://www.youtube.com)
2 [http://www.ted.com/](http://www.ted.com/)
3 [http://videolectures.net/](http://videolectures.net/)
4 [http://www.slideshare.net/](http://www.slideshare.net/)
where $key_1$ is the keyword string (topic label), and $P(key_1|resource)$ is the language model matching score, the likelihood of the keyword given the resource content. $P(resource)$ in this formula is the resource topic prior probability:

$$P_{prior}(resource) = P(z_{key_1}|resource)$$

which is the topic $z_{key_1}$ inference probability score given the resource content. Unlike the keyword string based language model, $P(z_{key_1}|resource)$ employed all the possible terms in the resource to “vote” for the topic level match as the resource prior. Then, for Figure 2 case, the second Wikipedia page, about the music album, will get a very low topic score, resulting in a low rank.

Ideally, we should also consider search engine rank as a kind of prior. However, unlike other meta-search problems, the ranking lists of some cyberlearning types in this research were totally disjointed. For example, for videos we sent queries to YouTube, TED, and Videolecture. Their indexes are almost disjoint, a feature that some meta-search ranking fusion algorithms do not appreciate, e.g., Borda’s method and Markov chain methods (Dwork, Kumar, Naor & Sivakumar, 2001).

**Topic Modeling on Publications**

Blei et al., (2003) proposed Latent Dirichlet Allocation (LDA) as a promising unsupervised topic modeling algorithm. LDA employs a generative probabilistic model in the hierarchical Bayesian framework, and extends PLSI by introducing a Dirichlet prior on $\theta$. As a conjugate prior for the multinomial topic distribution, the Dirichlet distribution assumption has some advantages, including simplification of the problem. The probability density of a $T$-dimensional Dirichlet distribution over the multinomial distribution $p = (p_1, p_2 \ldots, p_T)$, where $\sum \alpha_j = 1$, is defined by:

$$Dir(\alpha_1, \alpha_2 \ldots \alpha_T) = \frac{\Gamma(\sum \alpha_j)}{\prod \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j-1}$$

where $\alpha_1, \alpha_2 \ldots, \alpha_T$ are parameters of the Dirichlet distribution. These parameters can be simplified to a single value $\alpha_{LDA}$, the value of which is dependent on the number of topics.
However, one limitation of LDA is the challenge of interpreting and evaluating the statistical topics. For example, in this research, it is hard to assign a keyword to each statistical topic automatically. In addition, arbitrary numbers of topic may not be appropriate for this study because, while some topics may be very sparse (covering several keywords), others may focus only on quite detailed knowledge of the same scientific topic (covering part of a keyword). These limitations led us to use a supervised or semi-supervised topic modeling algorithm, one stem from LDA, which employs existing scientific keywords as the topic labels.

In this research, we assume that each (author-assigned) scientific keyword is a topic label and that each scientific publication is a mixture of its author-assigned topics (keywords). As a result, both topic labels and topic numbers (the total number of keywords in the metadata repository) are given. The labeled LDA (LLDA) (Ramage, Hall, Nallapati, & Manning, 2009) was used in training the labeled topic model. Unlike the LDA method, LLDA is a supervised topic modeling algorithm that assumes the availability of topic labels (keywords) and the characterization of each topic by a multinomial distribution $\beta_{key_i}$, over all vocabulary words. For example, the following table is an example of the keyword–word topic probability:

<table>
<thead>
<tr>
<th>Search engine</th>
<th>Semantic web</th>
<th>Directed graph</th>
<th>Image retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>google</td>
<td>0.0173</td>
<td>0.0339</td>
<td>0.0151</td>
</tr>
<tr>
<td>log</td>
<td>0.0116</td>
<td>0.0257</td>
<td>0.0060</td>
</tr>
<tr>
<td>site</td>
<td>0.0037</td>
<td>0.0109</td>
<td>0.0046</td>
</tr>
<tr>
<td>visit</td>
<td>0.0034</td>
<td>0.0102</td>
<td>0.0039</td>
</tr>
<tr>
<td>page</td>
<td>0.0011</td>
<td>0.0064</td>
<td>0.0036</td>
</tr>
<tr>
<td>focused</td>
<td>0.0010</td>
<td>0.0061</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

During the Bayesian generative topic modeling process, each word ($w$) in a publication is chosen from a word distribution associated with one of that paper’s labels (keywords). The word is picked in proportion to the publication’s preference towards the associated label $\theta_{\text{paper,key}_i}$ and the label’s preference for the word $\beta_{\text{key}_i,w}$. Figure 3 visualizes the LLDA generative process. For each topic (keyword) $\text{key}_k$, one draws a multinomial distribution $\beta_{\text{key}_k}$ from the symmetric Dirichlet prior $\eta$. Then, for each publication, one builds a label set $\Lambda$ paper for the deterministic prior $\Phi$. Finally, one selects a multinomial distribution $\theta_{\text{paper}}$ over the labels $\Lambda$ paper from Dirichlet prior $\alpha$.

![Figure 3. LLDA Algorithm](image)

**Associating Resources with Publications**

For scientific publication metadata resource generation, the most straightforward method is to search for the paper title (exact match) in different search engines. However, this method has two major limitations. First, given a relatively long publication title, if we use exact string match, search engines will find very few results. Based on our experiment on 70 publications, on average only 0.35 resources were retrieved based on the queries in Table 1, replacing [Keyword] with [Paper Title]. We also found that the quality of resources was not good. A large portion consisted of publisher or digital library access pages for the target publication, which do not help scholars and students understand the paper. Second, the
cost of this method is quite high. Given the very large potential publication collection, we needed to send multiple queries for each publication. For example, based on Table 1, for each keyword or publication, we needed to send four queries to Google. However, most search engines restrict the number of automatic visits, which makes this method inefficient.

In this research we used a more economical method to cope with this problem. We assumed each publication was composed of a list of topics, with each topic represented by an author-assigned keyword. We then used the results from the previous section to estimate the publication resources. As Figure 4 shows, for each resource type, we first aggregated all resources retrieved using the publication keywords and then used a ranking algorithm to identify the most important resources based on publication content. Such resources are highly likely to be relevant to the publication content and could assist students or scholars to capture the essence of the publication. For this method, we assumed that improving the understanding of scholars or students on the topics (keywords) of the paper would eventually help them to understand the paper itself.

Figure 4. Publication resources generation

For all publication resources, we can use the language model to rank all the resources. Specifically, we used the language model with Dirichlet smoothing (Zhai & Lafferty, 2001) to rank the resource, the likelihood of paper content given a resource:

$$\log P(\text{paper}|\text{resource}) = \sum_{w_i \in \text{paper}} \log \frac{c(w_i; \text{resource}) + \mu P(w_i|C)}{|\text{resource}| + \mu}$$

Where the ranking score is the sum of log likelihood of each word $w_i \in \text{paper}$. The smoothing function used word resource frequency, $c(w_i; \text{resource})$, and the resource collection probability, $P(w_i|C)$. In practice, the paper can be represented by either publication title or abstract. When we compare this method with the task introduced previously in the resource collection section -- keyword relevance computation, we have stronger confidence in the language model usage, because keyword strings are usually very short, only 2 or 3 tokens, but paper title and abstract provide much richer content, and the ranking performance should be much better. However, we also want to test and compare the topic prior performance in this study, so we define

$$p_{prior}(\text{resource}) = P(z_{\text{paper}}|\text{resource})$$

where $z_{\text{paper}}$ is a distribution of all the possible topics (mixture) of paper. Unlike LDA-based topic inference, we don’t have to project the paper onto the entire possible topic space, as the paper topic labels (publication keywords) are already available. Also, we assume the keyword is not equally important
for each publication; that is, some keywords are more specific than others, and may deserve higher “importance” scores. We used KF-IDF to calculate this importance, where:

\[
KF = \frac{C(key, paper)}{\log |paper|}
\]

\[
IDF_{key} = \log \frac{N}{C(key, paper)}
\]

As a variation on TFIDF, KF is the normalized keyword frequency in the paper’s title and abstract. IDF (inverse document frequency) has been used in IR as a measure of the general importance of a term in a collection. Similarly, we used IDF to assess keyword importance for our IR domain publication collection. In the IDF function, N is the total number of domain publications, and \( C(key, paper) \) is the number of publications with the target keyword, \( key \). We assume that if a keyword is rare in a collection (large IDF), this keyword could be more important for the paper. Then the resource (topic) prior can be characterized as:

\[
P(z_{paper}|resource) = \sum_{i=1}^{n} KFIDF_i \cdot P(z_{key}|resource)
\]

In this formula, the target paper has \( n \) keywords (topics), and the resource topic prior is the normalized sum of all the topic inference probability. The weight of each topic inference score is the KFIDF value.

Experiments

To test and compare different methods of cyberlearning resource metadata generation, we designed an experiment with information retrieval topics and publications. As we needed to judge whether a resource could effectively help scholars or students understand the essence of a scientific topic or publication, we invited graduate students with basic knowledge in information retrieval as volunteers for this experiment.

Unfortunately, some participants in this experiment did not have enough programming experience to provide judgments for source code resources. As a result, participants only evaluated five categories of resources in this experiment: Wikipedia pages, video lectures, tutorials, datasets, and slides.

Dataset and Topic Modeling

In this experiment, a total number of 20,799 information retrieval publications from 1965 to 2010 were used. We first used purposive sampling to identify 15 core conference proceedings and journals covering information retrieval, such as SIGIR, TOIS, CIKM, and others. Publications in these proceedings and journals were used as seed publications. Cited publications in these were then investigated to expand the corpus. If a paper was cited more than twice by these seed publications, we embodiment it into the test collection.

In the metadata repository, some publications did not have keyword metadata. To solve this problem, we first created a popular keyword (frequency > 3) list from the existing keywords in the test collection. We then searched for each keyword in the paper title and abstract by using greedy matching. For example, if “music information retrieval” was in the title, we didn’t use the keyword “information retrieval”. Matched keywords were used as “pseudo-keyword” metadata for the target publication if author-assigned keyword metadata was unavailable. The keyword and publication collections were used to calculate \( IDF_{keyword} \).

A total of 7,293 publications were sampled for LLDA topic modeling. Author-provided keywords were selected as topic labels. For instance, this paper has six author-assigned keywords. Thus, our LLDA training would have assumed that this paper is a multinomial distribution over these six topics. During preprocessing we also clustered similar keywords if the edit distance between them was very small (as in “k-
means” and “k means”) or if two keywords shared the same stemmed root (as in “web searches” and “web search”).

The resource topics may or may not focus on the scientific topics we extracted from the papers. For instance, Figure 2 shows a Wikipedia page about the topic “music.” In order to model the “noisy” topics (distributions), 4,668 web pages, labeled with ODP categories, were randomly sampled from the ODP (Open Directory Project) 5 database. The top page categories were used as the noisy topic labels: Arts, Games, Kids and Teens, Reference, Shopping, Business, Health, News, Regional, Society, Home, Recreation, and Sports. Two categories, Computers and Science, were not used for noisy modeling, as they are related to information retrieval research.

If a keyword appeared less than 10 times in the selected publications, we removed it from the training topic space. For publication content we first used tokenization to extract words from the title, abstract, and publication full text. If the word contained fewer than three characters, this word was removed. Snowball stemming was then employed to extract the root of the target word. We also removed the most frequent 100 stemmed words and words appearing less than three times in the training collection.

Finally, we trained an LLDA model with 605 topics (594 scientific keywords + 11 ODP categories). These topics were used to infer the resource topic distribution, \( P(z_{key}|resource) \). We then sampled 70 publications from SIGIR and CIKM conference proceedings for evaluation (this was the “publication task”). A total number of 401 keywords were contained in these publications. Finally, we sampled 45 keywords for the topic or keyword evaluation task (the “keyword task”).

Experimental Setting

As stated in the previous section, two tasks need to be performed in this experiment: a topic (keyword) evaluation task, and a publication evaluation task.

For topic evaluation, we sampled 45 keywords from the collection of 401. For each keyword, we designed a web evaluation page. Likewise, we designed an evaluation for each publication. On each evaluation page, the keyword or publication title and abstract were presented at the top, and the top cyberlearning resources that had been retrieved were offered with actual resource links. Users could access these resources via a pop-up window by clicking the hyperlink. Reading (or watching) and understanding these resources could be a time-consuming job for users, especially for the publication task, since participants needed to read and understand the topics of the paper first. As a result, we only offered users the top five resources for each resource type for each keyword page (hence, there were approximately 25 resources on each page), and the top three resources for each resource type for each publication page (hence, approximately 15 resources on each page). Each user was asked to evaluate ten publications and seven keywords. At this stage, we didn’t have a chance to find the best ranking method so we used the tentative method, language model without topic prior, for resource ranking for both keyword and publication tasks.

Seven graduate (masters or doctoral) students with basic knowledge in information retrieval participated in this evaluation. Each student can evaluate resource quality by using a dropdown menu. In this evaluation, students were requested to judge whether the cyberlearning resource could help them understand the essence of the scientific keyword or publication. Possible evaluations were “not relevant” (score of 0), “low relevance” (1), “good relevance” (2), and “high relevance” (3). Before performing this evaluation task, we trained participants with examples of helpful and unhelpful resources. The small number of participants is a major limitation for this research, and we will address it in the future work section.

\(^5\) http://www.dmoz.org/
Experimental Results

In this section, we report two kinds of experimental results based on analyzing students' judgments of relevance. First, we evaluate whether the automatically generated cyberlearning resource referential metadata helped students to better understand the essence of scientific topics and publications by calculating mean average precision (MAP) and mean reciprocal rank (MRR). MAP calculates the average percentage (precision) of the "good relevance" (2), and "high relevance" (3) in all the listed resources, and MRR measures the average rank of the first "good relevance" or "high relevance" in the ranking list. Higher MAP and MRR scores indicate that users are more likely to get the useful resources and to understand the essence of the topic and publication.

Second, we evaluate the resource-ranking algorithm(s) for both keyword and publication tasks. In this research, we used normalized discount cumulative gain (NDCG) (Järvelin & Kekäläinen, 2002) to validate the ranking algorithms. NDCG estimates the cumulative relevance gain the user receives by examining retrieval results up to a given rank on the list. NDCG is based on two assumptions: first, highly relevant documents are more valuable than marginally relevant documents (graded relevance); second, the lower the ranked position of a relevant document, the less valuable it is for the user. A ranked vector $V$ of results $[\text{label}(v_1), \text{label}(v_2), ..., \text{label}(v_n)]$ can be generated for each query $q$, where each item in the vector is a judgment of degree of relevance. In this research, 1 is not relevant and 4 is perfect relevance. With this vector, calculation of the discount cumulative gain (DCG) is possible:

$$ DCG@k(V) = \sum_{i=1}^{k} \frac{1}{\log_2(1+i)} (2^{\text{label}(v_i)} - 1) $$

The normalized DCG (NDCG) of $V$ is defined as the DCG vector divided by the ideal permutation of $V$. A perfect ranking algorithm returns $DCG@k(V) = \text{ideal DCG}@k(V)$ and NDCG score $= 1$. All NDCG calculations are then relative values on the interval 0 to 1.

Keyword Task Results

For the keyword task, the result of each kind of resource ranking result is presented in Table 3 and Figure 5. The best-performed number for each row is highlighted in the table. LM used language model and Dirichlet smoothing with keyword string as query. LLDA used topic inference score, $P(z_{\text{key}}, \text{resource})$, for ranking. LM + LLDA used topic inference score as the resource prior, and language model + Dirichlet smoothing for ranking.

Figure 5. Keyword Task resource quality chart (MAP, MRR, and nDCG)
Table 3
Resource quality for the keyword task

<table>
<thead>
<tr>
<th></th>
<th>LM</th>
<th>LLDA</th>
<th>LM+LLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>MAP</td>
<td>0.7576</td>
<td>0.7825</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.7820</td>
<td>0.8088</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.7639</td>
<td>0.8070</td>
</tr>
<tr>
<td>Video</td>
<td>MAP</td>
<td>0.7753</td>
<td>0.7361</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.8582</td>
<td>0.8594</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.7904</td>
<td>0.7510</td>
</tr>
<tr>
<td>Slides</td>
<td>MAP</td>
<td>0.7156</td>
<td>0.7667</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.7835</td>
<td>0.8199</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.7539</td>
<td>0.8012</td>
</tr>
<tr>
<td>Tutorial</td>
<td>MAP</td>
<td>0.7002</td>
<td>0.7410</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.7698</td>
<td>0.8120</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.7413</td>
<td>0.7822</td>
</tr>
<tr>
<td>Dataset</td>
<td>MAP</td>
<td>0.6243</td>
<td>0.6504</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.7959</td>
<td>0.8204</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.6667</td>
<td>0.6905</td>
</tr>
<tr>
<td>ALL</td>
<td>MAP</td>
<td>0.6771</td>
<td>0.7236</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.6613</td>
<td>0.7151</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.7300</td>
<td>0.8285</td>
</tr>
</tbody>
</table>

Overall, the average MAP across different resources was 0.6831 and the average MRR was 0.7372, which means most students in this evaluation believed that the cyberlearning resources (which were automatically generated via meta-search) could help them understand the keyword-based scientific topic, and that the quality of those highly ranked resources was good. Among different resource types, Wikipedia pages and video lectures performed better than datasets and slides, and datasets performed the worst.

Form a ranking perspective, NDCG suggests the ranking’s normalized discount cumulative gain, which is more sensitive to the degree of relevance (or resource usefulness). Clearly, considering topic prior will boost the ranking algorithm via prioritizing those useful resources based on user feedback.

We also found that, for wiki and video resources, LM + LLDA outperforms LLDA. However, for all other types of resources, LLDA is better than LM + LLDA. The main reason for this is the quality of the resource content. For wiki resources, we used Wikipedia dump, and the quality for each wiki page is very high. Similarly, all video results come from specific websites, such YouTube and Videolecture, which enable us to use regular expression to extract the video tags and description from HTML code. As a result, textual representation for a video resource is also accurate. However, for all other resource types, the results were mainly from Google, and we didn’t have an opportunity to identify the essential content from the HTML code because of the variations of web page structures. Subsequently, the content of HTML used for resource representation could be so noisy that the performance of the language model will be threatened. We will address this problem in the future work section.

Publication Task Results

Table 4 and Figure 6 shows the results of five different ranking methods for the publication task. LM Title (Title + Prior in Figure 6.) used the paper title as the query to search the content of the resource with language model combined with Dirichlet smoothing, and LM Abstract (Abstract + Prior in Figure 6.) used the paper abstract as query. LLDA (LLDA only) used the paper and topic match probability, $P(z_{\text{paper}}|\text{resource})$, for ranking proposed in the “Associating resources with publication” section, and we also considered KF-IDF weighting. Title + LLDA (Title) used topic match probability as resource prior probability and title based language model for ranking. Abstract + LLDA (Abstract) used topic prior with abstract based language model.
Overall, in terms of ranking, LM Abstract performs the best, and LLDA-based topic modeling as resource prior did not provide stable performance gain. LLDA alone worked worst. The main reason for this could be the quality of query. Compared with the keyword task, the publication task’s query length is much larger and query quality could be much better. For instance, the average length of a keyword query was 2.02 tokens, while the average title query was 8.86 tokens and the average abstract query was 232.75 tokens. It is clear that a long query contains much richer information (that is, a long query also contains the topic information itself), and the language model-based content match itself works well. This is the reason why LM Abstract also outperforms LM Title.

In terms of resource quality, the MAP and MRR in Table 4 illustrate that the performance of the publication task is lower than that of the keyword task, but students usually suggest that the highly ranked resources could help them understand the essence of the target publication. For instance, the MRR for all resources was 0.7187, which means students find the first or second recommended resource on the ranking list is helpful or very helpful. Similarly, datasets performed the worst while Wikipedia Pages, Video Lectures, and Slides are more useful for participants.

Table 4: Resource quality for the publication task

<table>
<thead>
<tr>
<th></th>
<th>LM Title</th>
<th>LM Abstract</th>
<th>LLDA</th>
<th>Title + LLDA</th>
<th>Abstract + LLDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>MAP</td>
<td>0.6042</td>
<td>0.6250</td>
<td>0.5125</td>
<td>0.5375</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.8891</td>
<td>0.9033</td>
<td>0.8186</td>
<td>0.8643</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.6083</td>
<td>0.6375</td>
<td>0.5208</td>
<td>0.5458</td>
</tr>
<tr>
<td>Video</td>
<td>MAP</td>
<td>0.7019</td>
<td>0.6704</td>
<td>0.5926</td>
<td>0.7185</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.8733</td>
<td>0.8546</td>
<td>0.7668</td>
<td>0.8896</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.7111</td>
<td>0.6778</td>
<td>0.6667</td>
<td>0.7185</td>
</tr>
<tr>
<td>Slides</td>
<td>MAP</td>
<td>0.6156</td>
<td>0.6361</td>
<td>0.5001</td>
<td>0.6190</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.8313</td>
<td>0.8427</td>
<td>0.5982</td>
<td>0.8513</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.6122</td>
<td>0.6463</td>
<td>0.5001</td>
<td>0.6259</td>
</tr>
<tr>
<td>Tutorial</td>
<td>MAP</td>
<td>0.6007</td>
<td>0.6163</td>
<td>0.4201</td>
<td>0.5932</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.8393</td>
<td>0.8374</td>
<td>0.5465</td>
<td>0.8128</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.6215</td>
<td>0.6319</td>
<td>0.6423</td>
<td>0.6215</td>
</tr>
<tr>
<td>Dataset</td>
<td>MAP</td>
<td>0.5474</td>
<td>0.5779</td>
<td>0.4116</td>
<td>0.5166</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.8045</td>
<td>0.7879</td>
<td>0.6198</td>
<td>0.7722</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.5725</td>
<td>0.6014</td>
<td>0.4659</td>
<td>0.5362</td>
</tr>
<tr>
<td>ALL</td>
<td>MAP</td>
<td>0.5991</td>
<td>0.5904</td>
<td>0.4694</td>
<td>0.5557</td>
</tr>
<tr>
<td></td>
<td>NDCG</td>
<td>0.6772</td>
<td>0.6607</td>
<td>0.5507</td>
<td>0.6607</td>
</tr>
<tr>
<td></td>
<td>MRR</td>
<td>0.7172</td>
<td>0.7187</td>
<td>0.6631</td>
<td>0.6847</td>
</tr>
</tbody>
</table>
Conclusion and Future Work

This research focused on the problem of generating cyberlearning referential metadata for scientific keywords and publications. It was conceptualized as an information retrieval and use problem amenable to automation using meta-search methodologies.

Based on preliminary evaluation, we found that automatically generated cyberlearning resources can help students better understand the essence of the scientific topic and publication. Meanwhile, the cost of this approach is very low. The ranking algorithms based on sophisticated content match (language model) and topic modeling (LLDA) can also effectively enhance the quality of resource ranking. For the keyword task, given short keyword queries, topic modeling can effectively help students better locate the most informative resources and remove the noisy ones. The publication task, on the contrary, used paper title and abstract as queries, which quality is much better, and topic modeling in this task couldn’t provide stable performance gain.

We also found that some types of resources are more helpful than the others. For instance, most students favored Wikipedia pages, video lectures, and presentation slides, instead of datasets.

There are two major limitations for this paper. First, we only recruited seven graduate students for the evaluation, which is mainly because of the cost and requirements of the tasks. For instance, in order to participate in this task, students should understand the basic concepts in information retrieval, and they are required to read (or watch) a large number of publications and resources for evaluation. In the future, we should generalize this experiment to other scientific domains, while finding a larger number of participants.

Second, the quality of the resource text is still questionable. Right now, we find Wiki and video performance are better than other types, which may be because of the quality of text. In the future, we should use some other algorithms to identify the important content from the HTML code for all resource pages, which could further improve the ranking performance.

References


Understanding the EMR Error Control Practices among Gynecologic Physicians

Ritu Khare  
Drexel University College of Medicine  
риту.харэ@nih.gov

Yuan An  
Drexel University College of Information Science & Technology  
yuan.an@ischool.drexel.edu

Sandra Wolf  
Drexel University College of Medicine  
sandra.wolf@DrexelMed.edu

Paul Nyirjesy  
Drexel University College of Medicine  
paul.nyirjesy@DrexelMed.edu

Longjian Liu  
Drexel University School of Public Health  
l85@drexel.edu

Edgar Chou  
Drexel University College of Medicine  
edgar.chou@DrexelMed.edu

Abstract

An “EMR error” refers to any incorrect, incomplete, or inconsistent patient information entered into electronic medical records (EMRs). Currently, the administering clinicians “manually” resolve such errors. Designing automated error control algorithms is a significant, and yet under-explored, informatics problem. In this study, we assess the EMR error detection abilities of physicians, reveal their strategies, and draw implications for computational algorithm design. Focusing on gynecologic practice, we conducted an error simulation study by fabricating several “erroneous” patient visit notes. We presented these notes to 20 experienced gynecologists, and asked them to detect any errors. Despite devoting substantial time, the participants could detect <50% of the introduced errors. Nevertheless, the successful cases helped reveal the 5 kinds of automatable “triggers” that helped participants sense an error candidate. The participants were able to recognize these triggers because of their comprehensive gynecologic knowledge accumulated through experience and medical school training.

Keywords: algorithms, data errors, EMRs, physicians, user study

Motivation

Electronic medical records (EMRs) have revolutionized the accessibility, legibility, and decisive ability of patient health information. However, the unusable EMR interfaces, situated within a demanding clinical environment, make the process of data-entry very error-prone. Quite often, clinicians inadvertently make mistakes while documenting patient visits and diagnosis information, and thereby commit “EMR errors,” which include incomplete, inaccurate, or inconsistent information (Brown & Patterson, 2001; Phillips & Gong, 2009). EMR errors are expensive; not only do they lead to poor data quality, but also they have the potential to cause unsafe quality of care, and to hold the physicians liable for medical malpractice (Classen, Pestotnik, Evans, Lloyd, & Burke, 1997; Fichman, Kohli, & Krishnan, 2011). Given the gravity of the problem, informatics research should actively engage in developing computational error control algorithms to alert the physicians in real time, and minimize further medical errors (Redwood, Rajakumar, Hodson, & Coleman, 2011). Since the EMR errors are largely underexplored, it is important

Acknowledgements: We sincerely thank the two anonymous reviewers for providing valuable suggestions for revising this paper.  
This research is supported in part by the Drexel’s Jumpstart grant for health informatics. The user study was approved from the Institutional Review Board at Drexel University. We thank all the user study participants for sharing their wisdom and time.


Copyright is held by the authors.
to follow an "inside-out" approach to develop algorithms, i.e., to (i) first understand the existing error control mechanisms, and (ii) then design the algorithms according to the observed limitations, and opportunities. Given the limited error control functionality provided by the existing vendor-designed EMRs, clinicians resort to certain "manual" techniques to review, detect, and resolve the errors (Phillips & Gong, 2009). In this study, we take the first step toward algorithm development, and systematically investigate the manual error control practices. We assess the abilities of physicians to detect a variety of EMR errors, elucidate their strategies, and accordingly, derive implications for algorithm design.

To understand the existing error control practices, we conducted error simulation and user study on the physicians who are responsible for electronically documenting a wide range of patient problems and visit information. We selected the extremely vital medical field of gynecology because one of the key investigators of the study worked as a data scientist in a women’s health research team affiliated with the College of Medicine at Drexel University. As a result, the team had a close interaction with the gynecologic physicians who document a variety of information pertaining to yeast infections, bacterial vaginitis, menstrual cycle issues, pre-natal and post-natal complaints, regular gynecologic examination, etc. In the future, we plan to conduct similar studies in other medical areas to validate the findings of this gynecology-specific study.

We conducted the study in context of outpatient clinics wherein clinicians document the patient visits into the EMRs in an on-the-spot “narrative” manner. The documentation occurs under extreme time constraints, and hence such an unstructured documentation was conducive to a variety of data errors (George & Bernstein, 2009). To simulate the clinic environment, we fabricated 7 gynecologic visit scenarios, and developed the corresponding EMR patient visit notes; one such note is shown in the Figure 1. To simulate the error-prone nature of data-entry, we purposefully introduced 97 errors of 5 different kinds into the notes. We conducted a user study individually with 20 gynecologic physicians having extensive experience with EMR visit documentation. The participants were presented with the flawed notes, and were asked to identify any data errors. The error detection step was followed up with a debriefing discussion to reveal the error detection and resolution strategies adopted by the participants.

We find that the participants could detect only 49% of the inaccuracy and inconsistency errors from the notes, and only 36% of the omission errors from the notes. This clearly indicates the need for developing automated algorithms that not only save time, but also provide a more effective error control solution. While the task performance of participants was very limited, we find that the strategies adopted by the participants are very important in developing guidelines for designing computational algorithms. The debriefing discussion suggests that there are certain data triggers that naturally prompt participants to sense a potential error. In our user study with gynecologic narrative notes, the participants relied on five kinds of error detection triggers: detection of abnormal examination results; recall of generic clinical guidelines; detection of abnormal history events; observation of discrepant information; and identification of broken information links. While decoding their strategies, we find that participants not only have the natural language processing (NLP) abilities, but also an immense amount of intuitive domain knowledge accumulated through experience and medical school training. To simulate such behavior, in addition to sophisticated NLP techniques, the algorithms should incorporate a wide range of federally established free resources for clinical guidelines, controlled vocabularies, drugs, diseases, drug indications, gynecologic best practices, drug interactions, etc. We briefly provide the linkages between the triggers, and the relevant trustworthy knowledge sources. The key contribution of this study is that, as a pre-step to design EMR error control algorithms, we explore an untapped knowledge source, i.e., the physicians, and learn algorithm design lessons from their abilities and behaviors. We plan to use the results of this study in implementing customized algorithms for the narrative EMR data specific to gynecologic patients.

The remainder of the paper is organized in the following manner. We first provide a background on patient visit notes and the typical errors associated with them. We then describe the results of the user study conducted with the gynecologic physicians, discuss the related literature, and conclude the paper.

---

In a typical outpatient clinic, the affiliated physicians administer several patient visits on a daily basis. For each visit, the associated provider is solely responsible for documenting complete visit information into the EMRs. Despite the provision to enter structured information into EMRs, providers find it far more efficient to enter impromptu narrative notes during patient visits (Doğan et al., 2010). Such a “patient visit note” should ideally record complete information associated with the visit; e.g., a typical gynecologic note is organized into the following 19 “sections” in that order: (1) Reason for Visit, (2) Chief Complaint, (3) History of Present Illness (HPI), (4) Allergies, (5) Current Medicines, (6) Active Problems, (7) Past Medical History (PMH), (8) Past Surgical History (PSH), (9) Family History, (10) Personal or Social History, (11) Gynecologic History, (12) Obstetric History, (13) Review of Systems, (14) Vital Signs, (15) Physical Examination (PE), (16) Assessment, (17) Tests, (18) Plan, and (19) Orders. Figure 1 shows a gynecologic note for a fabricated annual patient visit.

Although such notes capture rich information about a visit, their free-text nature, combined with the demanding clinical settings, often causes the administering clinician to commit errors. For instance, some of the errors in the above note are: (i) the physical examination section is not detailed enough, (ii) it is not specified who in the patient’s family had been diagnosed with hypertension and diabetes, and (iii) 5 out of the 19 sections are missing from the note. While appearing naïve and harmless, such entry-level errors often advance into more serious forms such as medication and prescription errors (J. C. Pham et al., 2012; Wetterneck et al., 2011). From the context of these narrative sectioned visit notes, we classify the entry-level errors into 5 broad categories: (a) inconsistent information, wherein the information presented between any two or more sections is contradictory to each other; (b) incorrect information, wherein the information is incorrect with respect to the scenario presented in the note, and to the clinical guidelines; (c) incomplete information, wherein certain essential information is omitted from the note; (d) missing section, wherein an entire required section is missing from the note; (e) miscellaneous errors,
such as placement of information into an inappropriate section, or usage of an un-established acronym\textsuperscript{2} in the note. An example of each type of error is highlighted in the Figure 1 and is elaborated in the Table 1.

Table 1
\textit{The 5 Kinds of EMR Errors}

<table>
<thead>
<tr>
<th>Kinds of Errors</th>
<th>Error Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Inconsistent</td>
<td>The \textit{reason for visit} conflicts with the active problem list</td>
</tr>
<tr>
<td>Information</td>
<td></td>
</tr>
<tr>
<td>b. Incorrect</td>
<td>The prescribed dosage for prenatal vitamins is 1 capsule (instead</td>
</tr>
<tr>
<td>Information</td>
<td>of 2 capsules) daily.</td>
</tr>
<tr>
<td>c. Incomplete</td>
<td>BMI suggests that the patient is obese, and yet the corresponding</td>
</tr>
<tr>
<td>Information</td>
<td>diagnosis and plan are omitted from the note</td>
</tr>
<tr>
<td>d. Missing Section</td>
<td>Certain sections, like \textit{chief complaints}, personal history, tests, etc.</td>
</tr>
<tr>
<td></td>
<td>are missing from the note</td>
</tr>
<tr>
<td>e. Miscellaneous</td>
<td>“supervision of normal first pregnancy” is placed in an inappropriate</td>
</tr>
<tr>
<td></td>
<td>section</td>
</tr>
</tbody>
</table>

It should be noted that all the errors described above can be perceived by carefully reviewing the note, some other kinds of errors such as, the patient forgetting to mention about an active medication, are beyond the scope of this work. In this study, we investigate whether physicians can detect the aforementioned 5 kinds of errors from patient visit notes. In addition, we reveal their strategies, and derive implications for algorithm design.

\textbf{The Study with Gynecologic Physicians}

The objective of this research is to study the manual error control practices among physicians. Through this study, we anticipate to learn from the physicians, and accordingly design computational error detection algorithms. To accomplish this, we conducted a user study with 20 gynecologic physicians, and presented them with several erroneous notes belonging to different hypothetical patients. The participants were asked to carefully audit the notes, and detect the errors. Each session with a participant was followed up with an open-ended debriefing interview. As a result of this study, we accomplished the following goals: (i) assess the participants’ ability to detect and resolve errors, (ii) explicate their intuitive strategies, and (iii) infer guidelines for algorithm design.

\textbf{User Study Design}

We recruited 20 gynecologic physicians (11 females, 9 males) working with the Department of Obstetrics and Gynecology in the Drexel University College of Medicine. Each participant had extensive documentation experience with the Allscripts\textsuperscript{3} EMR deployed into various affiliated clinics. The experience distribution of participants is shown in the Figure 2. Since EMRs were introduced to the facilities in 2008, the participants had at most 4 years of experience until the commencement of this study.

\textsuperscript{2} http://www.tabers.com/tabersonline/ub
\textsuperscript{3} http://www.allscripts.com
For the conduction of this study, we fabricated 7 visit notes belonging to different hypothetical gynecologic patients, and introduced several errors in each note. These erroneous notes were designed after discussion with the clinical investigators of this study, which had more than 20 years of experience with patient visit documentation. Though each introduced error is absolutely fabricated, it is largely inspired by real-world clinical documentation malpractices. The introduced errors comprise our “gold standard” list to be used for evaluation. We introduced a large number (total: 97) of errors in order to increase the odds of errors being detected by the participants, and thereby to increase the odds of learning about the manual strategies. Figure 3 shows the error distribution; as apparent, we introduced a large number of missing section (54) and incomplete information (25) errors because missing information is much more prevalent than other kinds of errors in practice (Botsis, Hartvigsen, Chen, & Weng, 2010; George & Bernstein, 2009; Smith, Banner, Lozano, Olney, & Friedman, 2009).

The study was conducted individually with each participant. During a typical session, the investigator had an in-person meeting with the participant; the participant was briefly introduced to the study, and was demonstrated some examples form the various kinds of errors. Each session was divided into two stages:

I. Analysis Stage: During this stage, the participant was presented with the paper prototypes of the patient notes. For each note, the participant was asked to carefully study the note, detect any data error(s), and document/annotate them on the same sheet of paper. In order not to overwhelm the participants, we did not ask them to categorize the errors.

II. De-briefing Stage: During this stage, the participant was asked of certain follow-up questions regarding the detected errors, e.g., what makes you conclude that certain data are erroneous? what in your medical training allowed you to detect this error? what measures would you take to resolve a certain error? why do you think these errors occur? The responses were transcribed for further analysis, and were synthesized to reveal their personal strategies and thought processes.

Throughout the sessions, we did not impose any time restrictions, and facilitated the participants to perform in an uninhibited manner.

Results and Findings

The Note Analysis Stage. During the analysis stage, each participant spent, on an average, 32.6 minutes to review the 7 visit notes, and detect the embedded errors. Interestingly, each detected error could be mapped to an item from the gold standard list of errors, leading to a perfect error precision by each participant. This finding stands in contrast to the existing EMR alert systems that often report false positives. To assess the participants’ detective abilities, we compute the error recall, which is the proportion of the errors detected by the participants with respect to the gold standard list of errors. To organize the results, we categorize the 5 kinds of errors into two bins based on the extent of physician liability: (i) hi-liability errors: This bin includes the errors that potentially affect the quality of patient care, and hence lead to higher liability to the concerned physician. This includes the incorrect and inconsistent information errors; (ii) mod-liability errors: This bin includes the errors that potentially affect the quality of data, and hence are relatively lower liability catalysts. This includes missing sections, missing information, and miscellaneous errors. It should be noted that the classification of an error kind into hi-liability and mod-liability is not absolute, and would ideally depend on the entire medical history of the patient. The graph in the Figure 4 shows the error recall performance (y-axis) of the participants (x-axis). On an average, the hi-liability and the mod-liability performances are 0.49 and 0.36, respectively. Based on the
within subject t-test, there is a statistically significant ($p<=0.05$) difference between the performances of the participants for the two categories of errors.

![Error Recall Performance Graph](image)

**Figure 4. Error Recall Performance**

We also measure the association of the task performance with the task duration, and the number of years of experience. The Pearson correlation coefficients are shown in the Table 2; these correlations are not statistically significant at the 0.05 level. Nonetheless, the best error recall performance (70%) was delivered by participant P5, who had 4 years of experience, and who spent 53 minutes auditing the notes. The lowest recall (17%) was achieved by participant P10, who had 1 year of experience, and who only spent 19 minutes analyzing the notes.

<table>
<thead>
<tr>
<th></th>
<th>Time spent</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi-liability</td>
<td>0.29</td>
<td>0.43</td>
</tr>
<tr>
<td>Mod-liability</td>
<td>0.1</td>
<td>0.19</td>
</tr>
</tbody>
</table>

**The Debriefing Stage.** The debriefing session lasted for an average 13.3 minutes per participant. During this stage, we encouraged the participants to think aloud about the errors and the adopted detection strategies. The results for this stage are also based on our observations during the analysis stage. All the participants were very confident of their performances during the note analysis stage, and were very vocal about their experiences. We discussed and attempted to uncover their abilities and strategies. In particular, we were interested in the provenance of detective abilities, the triggers that caused them to “sense” (Aron, Dutta, Janakiraman, & Pathak, 2011) the errors (See Table 3), and their opinions on the causes of EMR errors.

_How do you gain the ability to detect errors?_ Upon being asked about the provenance of their abilities, 6 participants attributed to the field experience of writing EMR notes in clinical settings; 4 attributed to their training in the medical school; 5 clearly mentioned that while the mod-liability errors can be detected by someone fresh out of medical school, the hi-liability errors could only be detected by someone who has extensive experience with on-the-spot clinical documentation. The remaining 5 participants attributed the origin of their abilities to both academic training and field experience.

_What are the triggers for detecting mod-liability errors?_ In context of the missing section errors, the participants unanimously mentioned that the format of a note is wired into their brains, and hence it is very easy for them to spot any missing sections. For example, the first note element a physician looks at is the chief complaint section, and it is easy to identify if this section is missing.
Identifying *incomplete or omitted information*, however, requires more attentive analysis of the note. We observed two different kinds of triggers that prompted the participants to detect such errors.

- **Detection of abnormal history events**: Participants could detect the *incomplete information* errors if a partially written abnormal history event caught their attention. For instance, the note in the Figure 1 shows a history of abnormal pap smear in the *gynecologic history* section; however, to completely qualify this event, more information, e.g., the diagnosis date, should be documented. As another example, the same note demonstrates a history of hypertension in the *family history* section, but does not state which family member suffered from hypertension.

- **General Information Recall**: The participants were able to detect omitted information when they could recall some general and mandatory checks associated with annual visits, e.g., the Centers for Disease Control and Prevention\(^4\) stipulate that HIV screening should be offered on an yearly basis, and if the participant recalled this information, he/she could verify whether HIV screening was present in at least one of the sections.

**What are the triggers for detecting hi-liability errors?** With regard to the *high-liability errors*, we observed the following triggers that led the participants to detect the *incorrect and inconsistent information*.

a. **Observation of discrepant information between two sections**: Certain errors could be detected as soon as the participants noticed a clear mismatch between the same kinds of information from different sections. For instance, in one of the notes, the strength of the drug “Diflucan” is mentioned as 200mg in the *Plan* section, and the strength of the drug “Fluconazole” is mentioned as 20mg in the *Orders* section. Both the sections refer to the same drug because Diflucan is a market brand name for the ingredient Fluconazole. The strength information in the *Orders* section is incorrect because the tablet form of this drug only comes in 150 or 200mg strength. In our user study, only 5 participants detected and resolved this error, 1 participant identified the error but couldn’t resolve, and the rest either couldn’t figure out the link between the two drugs, or simply missed out due to lack of attention.

b. **Detection of abnormal results**: Certain errors could be detected as soon as the participants could detect some abnormal, and yet unaccounted for, numerical results. For instance, in one of the notes, the blood pressure of the patient was “150/80” that suggests systolic stage 1 hypertension and diastolic borderline prehypertension, and yet this abnormal result was not alerted in the note. In our user study, 4 participants highlighted this error and believed that the corresponding plan and assessment should be created; and 3 participants detected the error but weren’t sure if it was a typing error, or an omission error by the physician.

c. **Identification of broken information links**: In a high-quality note, there is a story-like logical flow from across sections, e.g., the *reason for visit* is investigated in the *physical examination section*, and the results of the examination are diagnosed in the *assessment* section, and the appropriate recommendation is documented in the *plan* section. Certain errors were detected once any “broken information link” across multiple sections was observed. For example, in the note in the Figure 5, the results of the wet mount, described in the *physical examination* section, suggest that the patient has yeast infections and bacterial vaginitis. Although this information is acknowledged in the *plan* section, it is omitted from the *assessment* section. The assessment is rather mentioned as trichomoniasis even though there is no corresponding indication in the results. In our user study, 8 participants were confident that trichomoniasis has been accidentally written in lieu of bacterial vaginitis, 5 participants were confused whether the information in *physical exam* is incorrect, or the *assessment* has been entered incorrectly.

\(^4\) [http://www.cdc.gov](http://www.cdc.gov)
Why do the errors occur? The participants also presented their views on the causes of errors. A majority (17 participants) commented that physicians are primarily responsible for the errors, and that they should improve their documentation behavior. The participants offered the following reasons and suggestions: (i) since most providers write for themselves, they make a lot of assumptions, leading to a poor quality note. On the other hand, the physicians who share patients, such as residents in training, write higher quality notes, (ii) clinicians should ask more questions of the patient to ensure complete information, and (iii) whenever possible, clinicians should write in a list format instead of a narrative format, since lists are more likely to get correctly audited.

Only 5 participants attributed the errors to system design: (i) the Allscripts EMR propagates all the active problems through previous visits, creating a lot of inconsistent and obsolete information, (ii) the clinicians tend to write free-text notes because the structured interface is not user-friendly.

Implications for Error Control Algorithm Design

The error recall performance suggests that despite their expertise and experience, participants could not detect more than 50% of the errors. Only 3 participants identified more than 55% of the errors, and the performances for at least 5 participants was below 30%. These results clearly indicate that the existing manual strategies for error control are not foolproof, and it is imperative to replace them with effective computational algorithms. The participants delivered statistically better performance for the high-liability errors than the mod-liability errors. This further underlines the significance of learning from their expert abilities to minimize potential physician liability. None of the computed correlations between performance and experience/time were significant. So, we cannot draw any clear conclusion regarding learning-based error control algorithms.

Although the participants’ quantitative performance is not impressive, the results of the debriefing stage teach us important lessons on algorithm design. The results on the provenance of abilities suggest that the future algorithms should incorporate domain knowledge from a wide range of sources, and also be able to learn and infer from the contextual information in the EMR data. More importantly, the analysis of the triggers for error detection and the associated strategies provided us with some concrete implications on gynecologic error detection algorithms.
Table 3
Error Detection Trigger Categories and the Algorithm Design Implications

<table>
<thead>
<tr>
<th>Trigger Category</th>
<th>Associated Note Sections</th>
<th>Examples</th>
<th>Algorithm Design Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Information Recall</td>
<td>Physical Examination, Plan</td>
<td>If it is an annual visit, then all 19 sections should be present in the note. If it is an annual visit, then HIV screening should be performed and documented in the Physical Examination section. If the patient is over 60 years of age, then a health monitoring plan should be created and specified in the Plan section.</td>
<td>Computational Technique: Basic if-then rules, extraction of key information such as age, visit type, etc. Knowledge Sources: Clinical guidelines • Centers for Disease Control and Prevention • American Congress of Obstetricians and Gynecologists</td>
</tr>
<tr>
<td>Detection of Abnormal Results</td>
<td>Physical Examination, Vital Signs, Review of Systems, Plan</td>
<td>Any abnormal body mass index should be alerted in the Plan and Assessment sections. Any abnormal blood pressure should be alerted in the Plan and Assessment sections.</td>
<td>Computational Technique: Basic if-then rules, automatic extraction of examination results. Knowledge Sources: • Davis’s Laboratory and Diagnostic Tests • Agency for Healthcare Policy and Research • Archimedes 360 Medical Calculator</td>
</tr>
<tr>
<td>Detection of abnormal history event</td>
<td>History of Present Illness, Current Medicines, Past Medical History, Past Surgical History, Family History, Personal or Social History, Gynecologic History, Obstetric History</td>
<td>If an abnormal pap smear was observed, then the diagnosis date should be noted. For observation of white discharge, the duration and the odor should be specified. If herpes is noted as part of the gynecologic history, then the diagnosis date and the frequency of outbreaks should be specified. For the family history of breast cancer, the relationship of the family member should be specified. For the social history of smoking, the duration and frequency should be specified, and an appropriate counseling should be specified in the Plan section. For any current medications, the dosage, frequency, and administration route should be specified.</td>
<td>Computational Technique: Advanced if-then rules, extraction of abnormal events and their attributes from relevant sections, extraction of medications and their attributes. Knowledge Sources: the 5Ws of information gathering basics, Conceptual model for drugs, disease conditions, and habits. • UMLS RxNorm • DailyMed: Current Medication Information • MedlinePlus: Trusted Health Information for You</td>
</tr>
<tr>
<td>Observation of discrepant information between two sections</td>
<td>All drug-related sections, Reason for Visit, Active Problem List</td>
<td>The reason for visit should be consistent with the active problem list. The information on the same drugs should match across different sections, e.g., plan and orders Different drugs should be</td>
<td>Computational Technique: Comparison of problems, and medications across sections, Drug and Disease recognition Knowledge Sources: Controlled vocabulary for describing problems and drugs, linkages between drug ingredients, and brand names (Li, Khare, &amp; Lu, 2012), drug-drug interactions.</td>
</tr>
</tbody>
</table>
compared to verify any possible adverse interaction, e.g., in the Figure 5, Lithium (in Current Medications) and Metronizadole (in Plan), are pharmacologically incompatible. In our user study, only 5 participants detected this error.

<table>
<thead>
<tr>
<th>Identification of broken links across multiple sections</th>
<th>Physical Examination, Assessment, Plan, Reason for Visit, Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each abnormal result from Physical Examination, should be linked to a corresponding diagnosis in the Assessment section.</td>
<td>Each diagnosis item should have a corresponding item in the Plan section.</td>
</tr>
<tr>
<td>Each diagnosis item should have a corresponding item in the Plan section.</td>
<td>Each Plan item should optionally have at least corresponding order in the Order section.</td>
</tr>
</tbody>
</table>

Computational Techniques: Extraction of results, diagnosis, plan, order information, linking items from different sections, and discovering the missing links. Knowledge Sources: drug indication (Névéol & Lu, 2010), prescriptions, physical examination resources.

- DrugBank: Open Data Drug and Drug Target Database
- Unified Medical Language System
- FDA National Drug Code Directory
- Classification of Diseases, Functioning, and Disability
- RxDrugs
- Davis’s Drug Guide

The implications for algorithm design are summarized in the last column of the Table 3. The key step in algorithm design is to be able to programmatically fire the triggers that allowed participants to detect the errors. This requires a combination of computational NLP techniques, and a wide range of medical knowledge resources. For each trigger, some suggested computational techniques, and certain specific authenticated knowledge sources are provided in the table. While a plethora of NLP techniques have been proposed earlier, the use of existing knowledge sources to simulate physicians’ knowledge is still limited and under-explored. In the future, we intend to systematically use these findings, and build algorithms for each trigger, while integrating the relevant knowledge sources.

**Study Limitations**

The biggest limitation of this study is the demanding schedule of our participants. Besides seeing patients, clinicians engage in research, conduct clinical trials, and serve on multiple committees. Thus, the participants could only devote very limited time to this study. Therefore, the set of derived implications are by no means complete. Also, the study has an inherent bias; while the participants knew in advance that the notes contain errors, in reality, such assumptions are not made while reviewing the EMR. In addition, the frequency of errors (average 13) introduced in each note is not based on any empirical evidence due to the lack of related work. There is a possibility that some participants assumed the notes to contain fewer errors, and terminated their analysis upon finding certain number of errors.

**Related Literature**

There has been a considerable interest in medical errors in the past; medical errors occur due to a variety of reasons (Wagner & Hogan, 1995) including, un-captured handwritten changes made in the paper chart, electronic data-entry errors, patient-initiated changes in medications, changes made by external clinicians, etc. There has been a substantial literature discussion on the kinds of medical errors. There are two ways of perceiving medical errors: at the inception, and at the conclusion. The latter category includes duplicate order entry, pharmaceutical, adverse events, and medication errors (J. C. Pham et al., 2012; Wetterneck et al., 2011). In this work, we are interested in the errors at their inception, i.e., at the patient record entry level. At this level, several classification schemes are possible. Wagner et al. (Wagner & Hogan, 1995) describe two kinds of clinical errors, incompleteness, i.e., missing observations, and incorrectness, i.e., inaccuracy in recording information. Aron et al. (Aron et al., 2011)
classify the errors as procedural errors, which are not justifiable under any circumstances, and interpretive errors, which are qualified based on circumstances and other contextual information. Redwood et al. (Redwood et al., 2011) classify errors based on user intentions as unintended errors, e.g., accidentally typing 100 instead of 10, and unanticipated errors that occur when a user deliberately deviates from standard clinical guidelines. Botsis et al. (Botsis et al., 2010) classify errors as incomplete, inconsistent, and inaccurate, in the context of clinical narrative text. In this work, we build upon the existing taxonomy, and focus on a specific EMR artifact, the patient visit note. We do not take into account the user intention, and only focus on error as it appears on the document. We classify errors into 5 categories: inconsistent information, incomplete information, omitted information, missing sections, and miscellaneous errors.

While data entry errors are very prevalent, the existing EMR systems are very limited in catching the errors and alerting the users. The EMRs usually provide error control for structured data. For instance, the Allscripts EMR provides warnings on drug-drug interaction, allergies, and duplicate orders. However, the number of warnings is so high, and their relevance is so low that this often causes confusion and possibilities of more errors (Goldberg, Shubina, Niemierko, & Turchin, 2010). Also, the list of diseases to choose from is too lengthy, e.g., a filtered search on “diabetes” shows a list of 150 options.

There are few existing works that focus on designing algorithms for minimizing and controlling the clinical errors. Wilderman et al. (Wildeman et al., 2011) develop algorithms to minimize errors in clinical trial databases. The algorithm relies on validation rules, warning messages, range checks, and mandatory field checks to minimize errors on the data entry forms. Mitchel et al. (Mitchel et al., 2011) develop error control mechanism for electronic data capture system wherein only critical error-prone fields are targeted, and the validation rules are designed accordingly. As opposed to targeting the research databases as the previous two works, Goldberg et al. (Goldberg et al., 2010) target the EMRs, and design algorithms for detecting errors in the quantitative patient weight information. They develop two versions of algorithm: one that detects errors in real-time, and the other that works in a retrospective manner. To detect the possible outliers, and hence the erroneous entries, they use a combination of statistical techniques such as threshold analysis, change threshold analysis, difference from mean, etc. They evaluate the algorithm on 186 weight entries from real EMRs, and find that the real-time version is 81% accurate, and the retrospective version is 96% accurate as compared to expert judgment.

There are several limitations of the existing algorithms. First, they are designed for structured clinical data, and are hence inapplicable to a large amount of EMR data, which are narrative in nature. Second, they are largely based on the detection of abnormal results trigger adopted by physicians. However, the remaining 4 triggers (See Table 3) that we derived from our study are not yet incorporated. This makes the existing algorithms largely incomplete and incompetent for catching different kinds of data errors. Finally, these algorithms do not consider using any established medical knowledge resources, and are hence are less likely to simulate the abilities of the knowledgeable physicians.

In this study, we do not necessarily propose a complete error control algorithm, but we investigate the manual strategies adopted by expert physicians to detect and resolve errors, and learn several lessons on effective algorithm design. We focus on the gynecologic field of medicine, and understand the nature of errors associated with a specific EMR document, the patient visit note. We conduct a user study with the gynecologic physicians who have substantial experience with note documentation. We identify the specific triggers for error detection, the associated computational strategies, and the trustworthy knowledge sources to be incorporated in the future error control algorithms.

Conclusions

In the United States, medical errors kill more people than highway accidents every year (Kohn, Corrigan, & Donaldson, 1999). Although federally funded EMRs have been installed in several major hospital and clinic facilities, there is no evidence of decline in medical errors. In fact, EMRs themselves lead to a new family of errors, the “EMR errors” (Phillips & Gong, 2009; Thyvalikakath et al., 2012). Our ultimate goal is to design sophisticated computational algorithms to alert the physicians about the EMR errors in a real-time fashion.

In this paper, we have taken the first step to algorithm design, and have explored an untapped knowledge resource, i.e., the physicians. We have explored their abilities to detect EMR data errors, and have derived algorithm design implications from their intuitive knowledge and personal strategies. To accomplish this, we have conducted a user study with 20 gynecologic physicians, wherein we presented
them with prototypes of several erroneous EMR patient visit notes, and asked them to detect incomplete, inconsistent, or incorrect information errors. The error recall performance (<50%) of the participants indicated that the existing manual abilities are neither efficient nor guaranteed. However, an in-depth investigation of manual strategies helped us learn several lessons on error control algorithm design. The study helped reveal 5 kinds of triggers that help the physicians sense an error candidate. To identify the triggers, and to detect and resolve the specific errors, physicians rely on an implicit knowledge base of clinical guidelines, and gynecologic best practices. Such a knowledge base has been learned and accumulated through several years of field experience and medical school conditioning.

In comparison to the manual expert strategies, the existing automated algorithms only scratch the surface of error control. In the future, we plan to leverage the findings from this study and design customized algorithms for gynecologic notes. In particular, we would build on the identified triggers, and design algorithms accordingly. To simulate the narrative information extraction, several NLP algorithms exist to extract drug, disease and specific clinical information from texts (Doğan et al., 2010; Li et al., 2012; Névéol & Lu, 2010). To simulate the physicians’ knowledge in the head, we intend to utilize, integrate, and organize several available trustworthy knowledge sources hosted by the US Government.

References


Effects of User Identity Information On Key Answer Outcomes in Social Q&A

Abstract

Social Q&A (SQA) services have been growing in popularity among health information seekers. Even though research has paid much attention to a variety of characteristics of SQA services to investigate how people interact with each other for seeking and sharing information, the issues of identity and anonymity in these services that might relate to key user outcomes have been understudied. Such issues are especially important when dealing with stigmatized health conditions or sensitive health-related questions where choices are made about the revealing and concealing of identifying information in SQA environments. In the current study, we identified 110 stigmatized health questions from Yahoo! Answers that contained varied amounts and types of identity information corresponding to a framework developed in the study. We found that there are differences for providing personal contact information in one’s profile when relating identity information in user profiles to identity information in user questions. Questions with a high amount of demographic information in questions tend to receive slightly higher average number of responses and take shorter time to receive the best answer for stigmatized health questions.

Keywords: social Q&A, health question-answering, identity information, self-disclosure, anonymity

Introduction

The Internet has become an important tool for seeking health information, with a majority of people (74%) finding health information online as opposed to using family and friends (12%) or contacting physicians (4%) to get health information (Hesse et al., 2005). In addition, the Pew Internet Project estimated that 8 million Americans seek health related information online a day and around 75% of U.S. Internet users search for health information online (Fox, 2006, 2008b). As a key part of that Internet search, social Q&A (SQA) services such as Yahoo! Answers (http://answers.yahoo.com) or WikiAnswers have significantly increased in popularity over the past decade as an online information seeking methods where an asker’s information needs are formed by natural language questions posed to other users who can answer the question or even offer feedback on the given responses (Choi, Kitzie, & Shah, 2012). According to the Hitwise report, the U.S visits to SQA have increased 889 percent from 2006 to 2008 (Tatham, 2008).

Due to popularity of SQA, people have started utilizing SQA to seek health-related information and health has become one of the most popular topics in these services (Oh, 2012). With the popularity of seeking and sharing health information in SQA, previous and current research has paid much attention to various topics related to health information such as information quality (Kim et al., 2008; Oh et al.,

1 http://answers.yahoo.com
2 http://wiki.answers.com

Copyright is held by the authors.
2011; Stvilia et al., 2009), motivations (Oh, 2012), and systemic approaches (Smedberg, 2007, 2008). Yet, there is a lack of research on how users disclose their own personal information in SQA environments and how that might be related to the responses one receives. This seems especially important in this context because several health topics come with stigma that may make them embarrassing or otherwise threatening to discuss. When dealing with stigmatized health conditions or sensitive health-related questions, choices about the revealing and concealing of identifying information (e.g., name, photo, gender, height, weight, age, location, disease, symptom, etc.) are especially important as users attempt to provide enough information to get answers but also protect their own privacy.

Thus, the purpose of this current study is to focus on the degree to which a SQA user’s identity is disclosed and how that relates to various response characteristics in the Q&A process. In particular, the current study aims to address the following research questions:

RQ1: Does user information in a profile influence the amount and/or type of identity information in a question posted to SQA?
RQ2: How does amount and/or type of identity information in a question relate to the number of responses in SQA?
RQ3: How does amount and/or type of identity information in a user’s question relate to the time to get the question resolved in SQA?

Related work

Social Q&A

SQA is a web based question-answering service where an asker poses a question and others provide responses to the given question for satisfying an asker’s information needs. In other words, SAQ is a form of information retrieval where the users’ information need is specified in the form of a natural language questions, and the desired result is self-contained answer (Bian et al., 2008). Shah, Oh, and Oh (2009) argued that research on online SQA can be divided into two major groups—user-based and content-based. For user-based studies, some research has conducted content analysis and divided the roles of users in two ways – seekers and sloths (Gazan, 2007).

The results show that more active users in SQA environments tend to receive more responses than sloths. Other work has attempted to propose a method for automatic identification of authoritative actors based on the number of best answers to given questions that are provided by authoritative actors (Wang, 2008). For content-based studies, Kim, Oh, and Oh (2007) analyzed comments that were left upon the selection of best answers and evaluated them by the best-answer selection criteria – content value, cognitive value, socio-emotional value, information source value, extrinsic value, utility, and general statement. Shah, Oh, and Oh (2008) studied the understanding of various characteristics of user participation; the study identified users in two distinctive roles – consumer and contributor in social Q&A sites. A recent study by Shah, Radford, Connaway, Choi, and Kitzie (2012) investigated why information-seeking questions fail in SQA and developed a typology for explaining why these informational questions failed to get answers.

Health Question-answering in Online Environments

The Pew Internet Center’s Internet and American life project announced that approximately 75% of patients attempted to search health-related information online (Fox, 2008a), Additionally, around 40% of respondents in another study (Baker et al., 2003) reported that they use the Internet to look for advice or information about health or health care in 2001. Moreover, to investigate health information related content, a rich body of previous research has paid much attention to online health information quality (e.g., Berland et al., 2001; Bock et al., 2004, Donald, Lindenberg & Humphreys, 1998; Pastore, 2001, Zeng et al., 2004) while other research attempted to focus more on health-related information seekers or answerers.

For example, Gualtieri (2009) argued that some people intend to choose the Internet as their information resource over doctors in order for consultation on health issues, and adolescents sometimes prefer to find information and advice on stigmatized related health questions (e.g., pregnancy, sexuality, etc) from anonymous peers rather than close friends or family members (Suzuki & Calzo, 2004).
Moreover, Oh (2012) focused on how and why health answerers are motivated to answer those questions in order to share their information, knowledge, and experiences in SQA. That study found that the most influential factor for top answerers and health experts is altruism, whereas personal gain is the least influential factor for which top answerers and health experts share their knowledge, information, and experience with others. Unlike those previous studies, the research presented here focuses on identity information in user profile and user question when asking stigmatized health questions and also investigates how that information might be related to responses in SQA.

Identity Information for Self-disclosure

The existing research about identity information in online environments has primarily paid attention to identity sharing and privacy concerns in general (e.g., Gross & Acquisti, 2005; Krishnamurthy & Wills, 2008; Strater & Lipford, 2008; Stutzman, 2006). However, some previous research has examined Internet use and (sharing) identity information with regard to health-related questions. For example, Webber and Wilmot (2012) reported anonymous postings made on the Somazone website where people share and seek information and advice about sexual assault or sexual coercion. The report classified typical comments types – validation, interpretation, condemnation of poster, self-disclosure, personal advice, medical advice, legal advice, resources, and humor; furthermore, young users perceive the site as a safe place where they can disclose themselves to receive professional nonjudgmental answers. Zhang (in press) focused on college students to study how they perceive and use the social networking sites for health and wellness information. The study revealed that 21% of participants responded that they would use the general health-related websites (e.g., WebMD) for getting health and wellness information rather than their existing social ties on Facebook because they do not want to broadcast their personal information to known people. In addition, other research (see Frost & Massagli, 2008, 2009; Brubaker, Lustig, & Hayes, 2010; Wicks et al., 2010) also specifically investigated PatientsLikeMe4 in an attempt to study how users share their personal information and access to personal health data for acquiring relevant health information.

Anonymity

According to Marx (1999), anonymity is "one polar value of a broad dimension of identifiability versus nonidentifiability" (p.100) and identity knowledge is "an aspect of informational privacy" (p.100) that has multiple components such as legal name, locatability, pseudonyms linked or not linked to name/location, pattern knowledge, social categorization and symbols of eligibility/noneligibility. In computer-mediated communication (CMC), anonymity is often defined in two distinct ways (Qian & Scott, 2007): discursive anonymity concerns the withholding of one’s personal information (name, email, gender, location, etc.) and visual anonymity concerns the absence of visual presentation (pictures, video clips, personally-identifying images, etc.) of people.

Moreover, Caspie and Gorsky (2006) found people are sometimes being anonymous to deceive their identity online because of privacy concern (21%) and Azechi (2005) argued that lack of personal specification as a type of anonymity causes communication to become more information-oriented and less social among people.

While previous research argues that being either visually and/or discursively anonymous online tends to create a more information-oriented environment, it is still unclear how different types and amounts of anonymity or identity information might have effects on getting information that satisfies an asker’s needs online. Thus, in this research, we focus on investigating how types of identity information and amount of identity information relate to responses from other users in SQA. To do so, we did conduct a content analysis study as described in the next section.

---

4 http://www.patientlikeme.com/
Method

Data Collection from Yahoo! Answers

400 resolved questions were collected by the Yahoo! Search Application Programming Interface (API)\(^5\). We looked at questions and answers in stigmatized health topics from a health category in Yahoo! Answers. Such keywords – abortion, depression, drug addiction, and STD were used to extract 100 resolved questions for each stigmatized health topic from the database.

Tourangeau and Smith (1996) aggregated previous research that focused on sexual behaviors and other sensitive topics in order to compare three methods of collecting survey data for sensitive questions; the study included drug use (Aquilino, 1994; Aquilino & LoSciuto, 1990), HIV risk factors (Locke et al., 1992), and abortion (London & Williams, 1990; Mosher & Duffer 1994; Mott, 1985). Moreover, other research argues that depression is one of significant topics in health informational seeking behaviors among teens (Boldero & Fallon, 1995; Dubow et al., 1990; Puskar, Tusaie-Mumford, Sereika, & Lamb, 1999; Borzekowski & Rickert, 2002).

Identifying Questions within the Framework of Amount and Type of Identity Information

To gain a better understanding of how different amounts and/or type of identity information in questions influences responses, we developed a framework for classifying both amount and type of identity information. Identity information includes profile data that reveals who an asker is and/or language in questions related to one’s identity.

“Amount of identity information” can be divided into three levels: questions with no identity information, questions with medium identity information (2 identifiers), and questions with high identity information (4 or more identifiers). “Type of identity information” takes three forms for those questions containing identifiers: demographic only (e.g., age, gender, grade, location), medical only (e.g., diseases, symptoms, medical history), or both demographic and medical identity information (see Table 1).

Table 1
Example of each amount and type of identity information.

<table>
<thead>
<tr>
<th>Identity information amount</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>No information</td>
<td>Does it hurt to have an abortion?</td>
</tr>
<tr>
<td>Medium number</td>
<td>my daughter is 14 can she take the Morning after Pill?</td>
</tr>
<tr>
<td>High number</td>
<td>Is this being weird or depressed? Maybe, is it even normal?? During school I phase out a lot, can't concentrate...so I just sit there and have daydreams about all kinds of things (I'll make up stories about me, not having so many problems in life, being happy, finding a great husband, travelling, etc.) And when I get home I sleep a LOT. I go straight to my bed, because I want to dream some more. I usually kinda recap my day, but in a more &quot;creative&quot; way. I sleep for hours during the day and don't feel like eating or going out. But the thing is...I'm really really sad, and think about my death sometimes, what its like being hurt...and I have nightmares about the time I was raped. This wouldn't be normal would it, or what is it?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Identity information type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic only</td>
<td>How do you take a drug addict to the rehabilitation centre?</td>
</tr>
</tbody>
</table>

\(^5\) http://developer.yahoo.com/answers/
The boy is 20 years old and will not listen to his mother. how do we help him

okay my parents are getting separated but not a divorce, and my dad just told me tonight that he already bought a house and at the end of this month hes going to move in. im okay with it-i knew it would happen but i cant stand to see my parents in this condition-sad, depressed, etc. Now i will have to take responsibility to take care of my mother, and especially my little brother whos only 10. he wont understand. but what do i do? im just 14. i have an older sister who is 17, a senior and next year she will graduate...

how do you know your stressing? like idk if i have or have not been!But it do feel like it..My attitude sucks rite about now && i feel bi-polar(but im not) && i nned to know if im stressin or need help..My head hurts alot also && i also feel a lil depressed

I take valiums regularly to fight my insomnia. Am I considered as a drug addict? I have a sleeping sickness called insomnia. It's hard for me to fall asleep during night-time but I feel sleepy and lethargic during daytime. My friend told me about Valiums. He said that it can help my sleeping problems. I gave it a try and it worked for me. Now, I found myself relying on Valiums to achieve that well-deserved good sleep. Am I addicted to it?

What can I do to help my sister who is a drug addict? My sister is a drug addict. She has been for over 5 years. She was raped while she was "high" on drugs and now has a son. She takes her son with her when she goes and buys drugs. She lives with our parents. She also has now stolen over $12,000 from our Grandmother, and our Grandmother is not going to press charges. They think it is okay to sweep this under the rug. But I'm fed up. I want to call CPS on her or something. What can I do?

When i was 16 i got pregnant and had an abortion now some years have passed and i think i might be pregnant.? I get irregular periods like i always have and so ive taken birth control pills for them since i was 14... i went off for a year and started again and took em for a couple months but they were making me feel sick so i had to go for a lower dose but through all this me an my boyfreind continue to have sex no big deal but now im feeling sick when i smell certain things an my nipples are really sore an ive been sleeping alot.. the only thing i can compare this to is my first pregnancy which happened right after i stopped taking my pills but i only had one symptom which was dark spotting for a day none of this.. any help?

(Note that the highlighted in questions are identifiers that fit within the framework)
Using this framework, content analysis of 400 stigmatized health questions was conducted by the coder to identify certain questions that fit in the framework. In the results, 110 of the 400 questions (27.5%) were identified that contain the relevant amounts and types of identity information (see Table 2). These came from across the abortion (n=26), depression (n=39), drug addict (n=30), and STD subcategories (n=15).

Questions with either one or three pieces of identity information were excluded in order to make amount levels more distinct. Moreover, the study also manually collected user profile information for each asker such as user nickname, photo, personal contact information, and the ‘about me’ page. However, other user profile information such as ‘points’ and ‘levels’ (which encourage the users to participate in a variety of activities, e.g., selecting the best answer) were not included in this study since users are not able to show or hide that information for others.

Table 2
Number of each identity information amount and type.

<table>
<thead>
<tr>
<th>Identity information amount</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>No information</td>
<td>19</td>
</tr>
<tr>
<td>Medium number</td>
<td>40</td>
</tr>
<tr>
<td>High number</td>
<td>51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Identity information type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic only</td>
<td>19</td>
</tr>
<tr>
<td>Medical only</td>
<td>24</td>
</tr>
<tr>
<td>Both demographic and medical</td>
<td>48</td>
</tr>
</tbody>
</table>

Results

Relating User Profile Identity Information to User Question Identity Information

Users asking stigmatized health questions generally provide minimal identity information in their profile. More specifically, they primarily use ‘no name or pseudonym’ (n=69, 62.70%) for a user nickname, either ‘no photo’ or ‘fake photo’ (n=105, 95.45%), no ‘contact information’ (n=57, 51.80%), and no details on the ‘about me’ page (n=91, 82.70%). When relating identity information in user profiles to identity information in user questions, results show there are differences for providing personal contact information in one’s profile (see Table 3).

First, there are differences in the amount of personal identity information in user questions relative to profile contact information, \( \chi^2 = 6.707, \text{ df}=1, p<.1 \). More specifically, users with a high amount of personal information in questions are slightly more likely to provide contact information in their profile (53%); users with a medium amount of personal information in questions provide contact information in their profile slightly less often (48%); and users with no personal identity information in questions are even less often including contact information in their profiles (37%).

Moreover, there are differences in the type of personal identity information in user questions relative to profile contact information, \( \chi^2 = 21.339, \text{ df}=1, p<.05 \). More specifically, users who provide only demographic identity information in questions tend to less often provide contact information in their profile (37%); users who use medical identity information in their questions provide contact information in their profiles more often provide personal contact information (54%) and users who include both demographic and medical identity information in their questions provide contact information in their profile even more regularly (58%). Except for contact information, there are no statistically significant differences between
other types of identity information in user profiles and the amount or type of identity information in user questions.

**Relating User Question Identity Information to Number of Responses**

Although statistical comparisons do not reveal significant differences given some of the small Ns in certain categories, it is worth looking at the descriptive for general trends in the data. Users with a medium amount of information in questions tend to receive slightly higher average number of responses (mean= 11.02, S.D.=10.05); users with a high amount of identity information (mean= 9.40, S.D.=6.87) or no identity information (mean=8.47, S.D.=7.03) received slightly less average number of responses.

More specifically, users with a high amount of demographic identity information in questions without any specific medical identity information received the highest number of responses (mean=13.11, S.D.=9.73), trailed by users who provide a medium number of medical identity information (mean=12.36, S.D.=12.38).

**Table 3**

**Profile Information Relative to Identity Information Amount and Type in Questions**

<table>
<thead>
<tr>
<th>Information amount</th>
<th>User ID</th>
<th>Photo</th>
<th>Contact</th>
<th>About me</th>
</tr>
</thead>
<tbody>
<tr>
<td>No information</td>
<td>No name/pseudonym (n=12, 63.16%)</td>
<td>No photo (n=11, 57.69%)</td>
<td>No contact (n=12, 63.16%)</td>
<td>No about me information (n=17, 89.47%), About me information (n=2, 10.53%)</td>
</tr>
<tr>
<td>Partial name/full name (n=7, 36.84%)</td>
<td>Fake photo (n=7, 36.84%), Real photo (n=1, 5.26%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium amount</td>
<td>No name/pseudonym (n=24, 60.00%)</td>
<td>No photo (n=23, 57.50%)</td>
<td>No contact (n=21, 52.50%)</td>
<td>No about me information (n=14, 73.68%), About me information (n=5, 26.32%)</td>
</tr>
<tr>
<td>Partial name/full name (n=16, 40.00%)</td>
<td>Fake photo (n=14, 35.00%), Real photo (n=3, 7.5%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High amount</td>
<td>No name/pseudonym (n=33, 64.70%)</td>
<td>No photo (n=28, 54.00%)</td>
<td>No contact (n=24, 47.56%)</td>
<td>No about me information (n=42, 82.35%), About me information (n=9, 17.65%)</td>
</tr>
<tr>
<td>Partial name/full name (n=18, 35.30%)</td>
<td>Fake photo (n=22, 43.14%), Real photo (n=1, 1.96%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information type</td>
<td>User ID</td>
<td>Photo</td>
<td>Contact</td>
<td>About me</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------</td>
<td>-------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Demographic only</td>
<td>No name/pseudonym (n=11, 57.89%)</td>
<td>No photo (n=14, 73.68%)</td>
<td>No contact (n=12, 63.16%)</td>
<td>No about me information (n=14, 73.68%), About me information (n=5, 26.32%)</td>
</tr>
<tr>
<td>Partial name/full name (n=8, 42.11%)</td>
<td>Fake photo (n=5, 26.32%), Real photo (n=0, 0%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical only</td>
<td>No name/pseudonym (n=16, 66.67%)</td>
<td>No photo (n=12, 50.00%)</td>
<td>No contact (n=11, 45.83%)</td>
<td>No about me information (n=18, 75.00%), About me information (n=6, 25.00%)</td>
</tr>
<tr>
<td>Partial name/full name (n=8, 33.33%)</td>
<td>Fake photo (n=11, 45.83%), Real photo (n=1, 4.17%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both</td>
<td>No name/pseudonym (n=30, 62.50%)</td>
<td>No photo (n=25, 52.08%)</td>
<td>No contact (n=20, 41.67%)</td>
<td>No about me information (n=42, 87.50%), About me information (n=6, 12.50%)</td>
</tr>
<tr>
<td>Partial name/full name (n=18, 37.50%)</td>
<td>Fake photo (n=20, 41.67%), Real photo (n=3, 6.25%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>No name/pseudonym (n=69, 62.70%)</td>
<td>No photo (n=62, 56.40%)</td>
<td>No contact (n=57, 51.80%)</td>
<td>No about me information (n=91, 82.70%), About me information (n=19, 17.30%)</td>
</tr>
<tr>
<td>Partial name/full name (n=41, 37.30%)</td>
<td>Fake photo (n=43, 39.10%), Real photo (n=5, 4.5%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Moreover, in terms of information type, users with demographic identity information only received a higher numbers of responses (mean=12.42, S.D.=9.86) than users with only medical identity information (mean=10.58, S.D.=10.22) or with both demographic and medical identity information (mean=9.12, S.D.=6.64). The specific distribution of responses among different amounts and types of identity information that disclose an asker’s identity can be seen in Table 4.

Table 4
Average number of responses among different numbers and types of personal information in stigmatized health questions on Yahoo! Answers

<table>
<thead>
<tr>
<th>Information amount</th>
<th>Number</th>
<th>Average number of response</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>No information</td>
<td>19</td>
<td>8.47</td>
<td>7.03</td>
</tr>
<tr>
<td>Medium number</td>
<td>40</td>
<td>11.02</td>
<td>10.05</td>
</tr>
<tr>
<td>Demographic</td>
<td>10</td>
<td>11.80</td>
<td>10.46</td>
</tr>
<tr>
<td>Medical</td>
<td>14</td>
<td>12.36</td>
<td>12.38</td>
</tr>
<tr>
<td>Both</td>
<td>16</td>
<td>9.2</td>
<td>7.79</td>
</tr>
<tr>
<td>High number</td>
<td>51</td>
<td>9.40</td>
<td>6.87</td>
</tr>
<tr>
<td>Demographic</td>
<td>9</td>
<td>13.11</td>
<td>9.73</td>
</tr>
<tr>
<td>Medical</td>
<td>10</td>
<td>8.10</td>
<td>5.82</td>
</tr>
<tr>
<td>Both</td>
<td>32</td>
<td>8.79</td>
<td>6.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Information type</th>
<th>Number</th>
<th>Average number of response</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only Demographic</td>
<td>19</td>
<td>12.42</td>
<td>9.86</td>
</tr>
<tr>
<td>Only medical</td>
<td>24</td>
<td>10.58</td>
<td>10.22</td>
</tr>
<tr>
<td>Both</td>
<td>48</td>
<td>9.12</td>
<td>6.64</td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td>9.82</td>
<td>8.18</td>
</tr>
</tbody>
</table>

Relating User Question Identity Information to Best Answer Selection

In terms of time to best answer selection, users with no identity information in questions, although no significant differences given some of the small Ns in certain categories, generally took the shortest time (in hours) to receive the best answer for stigmatized health questions in Yahoo! Answers (mean=14.78, S.D.=11.02), followed by users with a medium amount of identity information (mean=16.51, S.D.=11.90) and users with a high amount of identity information (mean=17.88, S.D.=11.13).

However, the results show that users with a high number of demographic identity information seem to take the shortest time to select the best answer. Additionally, users with only demographics identity information took the shortest time to the best answer selection (mean=14.67, S.D.=14.31) than users with only medical identity information (mean=17.02, S.D.=8.51) or users with both demographic and medical identity information (mean=17.61, S.D.=10.43) as shown in Table 5.
Table 5
Average time to best answer selection among different numbers and types of personal information in stigmatized health questions on Yahoo! Answers

<table>
<thead>
<tr>
<th>Information amount</th>
<th>Number</th>
<th>Average time (hour) to best answer selection</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>No information</td>
<td>19</td>
<td>14.78</td>
<td>11.02</td>
</tr>
<tr>
<td>Medium number</td>
<td>40</td>
<td>16.51</td>
<td>11.90</td>
</tr>
<tr>
<td>Demographic</td>
<td>10</td>
<td>15.75</td>
<td>19.62</td>
</tr>
<tr>
<td>Medical</td>
<td>14</td>
<td>16.78</td>
<td>9.13</td>
</tr>
<tr>
<td>Both</td>
<td>16</td>
<td>16.75</td>
<td>8.02</td>
</tr>
<tr>
<td>High number</td>
<td>51</td>
<td>17.88</td>
<td>11.13</td>
</tr>
<tr>
<td>Demographic</td>
<td>9</td>
<td>13.47</td>
<td>4.97</td>
</tr>
<tr>
<td>Medical</td>
<td>10</td>
<td>17.35</td>
<td>8.02</td>
</tr>
<tr>
<td>Both</td>
<td>32</td>
<td>19.29</td>
<td>12.92</td>
</tr>
<tr>
<td>Information type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Only Demographic</td>
<td>19</td>
<td>14.67</td>
<td>14.31</td>
</tr>
<tr>
<td>Only medical</td>
<td>24</td>
<td>17.02</td>
<td>8.51</td>
</tr>
<tr>
<td>Both</td>
<td>48</td>
<td>17.61</td>
<td>10.43</td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td>16.85</td>
<td>11.35</td>
</tr>
</tbody>
</table>

Discussion

The online environment continues to play a vital role in health-related information seeking and sharing activities. For various stigmatized health issues, people prefer to go online to seek health information with anonymous peers rather than meet face-to-face with known people (Suzuki & Calzo, 2004); however, whether sharing more personal identity information online influences the responses one receives has been understudied.

The results indicate that users asking such stigmatized health-related questions in SQA provide minimal identity information in their profile. It seems that online Q&A users who pose these health-related questions would rather provide their personal information in questions as needed than disclose identity in the profile for all to view. It may signify that discursive and visual anonymity that withhold not only personal information but also visual presentation (Qian & Scott, 2007) in online environments is normative—allowing people to participate in a variety of online interactions (Scott & Choi, 2012).

However, the only identity information in one’s profile related to amount and type of identity information in questions is personal contact information. This finding revealed that users who do not intend to provide any specific identity information in a question are less likely to provide personal contact information in their profile, whereas users who willingly provide either demographic and/or medical identity information are more likely provide personal contact information. People may consider personal contact information as a less identifiable factor than other profile information (e.g., photo, name, etc) or even expect further communications between an asker and responders via email or instant messaging with regard to an asker’s health issues. Yet, we do not know whether people actually communicate between an asker and responders via their personal modes of communication for health questions.

Future studies should be done in order to analyze what motivates users include their personal contact information in user profile, or to investigate whether an asker and responders directly interact with each other via their personal contacts beyond question-answering processes in SQA.

Compared to face-to-face interactions, online Q&A environments as a form of computer-mediated communication (CMC) lack social context cues (Walther, 1996), which may hamper understandings of a sender’s messages. This indicates that an asker may need to provide more information such as demographic and/or medical information when asking health-related questions in online environments for a better understanding of the current health issues and conditions. However, the results show that users with a medium amount of identity information received a relatively higher number of responses than users with a high amount of identity information with regard to an asker’s stigmatized health question. This
signifies that providing excessive amount of identity information when asking a question in SQA would discourage responses rather foster interpretations regarding an asker’s information needs. This finding also reflects that one of significant attribute of failure of fact-based questions (Shah et al., 2012) is that questions are complex; thus, questions that are too complicated and/or overly broad or provide excessive identity information are less likely to get a response because of that complexity. Moreover, these findings indicate that providing more identity information when asking stigmatized health questions does not necessarily mean better or clearer information.

One of the implications from these findings is that an asker may revise his/her question by reducing identity information that may be unnecessary and include only major identifiable information that helps responders gain a better understanding of an asker’s current health issues and conditions. Another suggestion would be to redirect an asker to other online health services for seeking health information (e.g., PatientsLikeMe, WebMD⁶, etc.) where people may interact with more professional health and medical related news and information. Further study should focus on comparisons between SQA and other online health services or forums to investigate how similar stigmatized health questions have either similar and/or different outcomes that satisfy an asker’s information needs.

The study also found that users with demographic identity information in a question seem to receive relatively higher numbers of responses and have the shortest time to best answer selection as compared to those who provide medical identity information. This may signify that such demographic identity information is more likely objective and readable than medical information. It also suggests that how an asker describes his/her medical symptoms or conditions in stigmatized health questions is relatively subjective in that people may describe a similar symptom in various ways by their own terms and expressions. This may hamper better understandings of responders on what an asker intends to ask, which may contribute to fewer responses and longer delays in selecting a best answer. Such subjective and ambiguous forms of information provided by health-related information seekers may limit opportunities to receive more responses from others in SQA environments. This result also supports other findings that ambiguity is one of the significant attributes of failed questions, which impairs better interpretation of the questions (Shah et al., 2012).

Another interesting finding is that users with no identity information (in terms of information amount) have shorter time to best answer selection (mean=14.78, S.D.=11.02). The questions with no identity information may not ask specific personal health issues or conditions; instead, they are more likely to ask the general issues in stigmatized health areas. The following example illustrates questions with no specific personal identity information in stigmatized health topics:

“Does it hurt to have an abortion?”

This question appears to look for more general knowledge about an abortion rather than specific advice or opinion about an asker’s health issues, which means that an asker may select any response as the best answer if the response somewhat fulfills an asker’s information need. In addition, users with either a medium and/or a high amount of identity information received a relatively higher number of responses than users with no identity information. As one way of accounting for this, Shah (2011) argues that a delayed answer may use earlier posed responses to the given question for a more comprehensive answer for an asker’s information needs so that delayed answers could have the likelihood of being selected as the best answer.

**Limitations**

Even though the study presented here provided insights into the effects of identity information in a user profile and a user question on responses for stigmatized health-related topics, the study is limited in several ways. For example, to provide a distinction between different amounts of identity information, the study did not include questions with either one or three pieces of identity information (although such amounts can, of course, be found on these sites). We do not know exactly how those excluded amounts might be similar or different to the high/medium/low amounts in terms of their influence on the quantity of responses. In addition, because this exploratory study only analyzed 110 question-answer pairs, it is difficult to generalize from this small, non-random sample. For this reason, future study might focus on a

---

⁶ http://www.webmd.com/
larger dataset to determine whether there exist significant differences between different amounts of identity information in stigmatized health questions. This study is also limited in evaluating responses because it only focused on how amount and/or type of identity information in a question relates to the number of responses and the time to get the question resolved in SQA. It is still arguable whether more and faster answers are the better answers to the given question. Future study could focus on evaluating answer quality based on different amount and types of identity information in stigmatized health questions in SQA.

**Conclusion**

The current study investigated the degree to which an SQA user’s identity is disclosed and how that relates to various response characteristics in the question-answering process. Different amounts and types of identity information in a question have been studied regarding stigmatized health questions in SQA. This work may be a useful early step in understanding disclosure of identity information during question-answering processes for people seeking professional opinions and advice for such stigmatized health-related questions. Findings from the current study about identity information amount and type shed light on how an asker discloses his/her identity information for a better understanding of his/her health issues and conditions in stigmatized health questions in SQA. An appropriate amount of objective and less ambiguous identity information helps responders understand what an asker looks for and encourages them to provide a response to his/her question.

This study also makes a contribution to anonymity and identity information in online environments by exploring how identifiable information that discloses an asker’s identity relates to receiving responses from others and/or getting his/her information needs resolved in SQA. Previous research argues that lack of personal specification causes less social and more information oriented communication among people online (see Azechi, 2005). Yet, the findings in this study suggest that disclosing appropriate amounts and types of identity information should be considered when posing a question in order for better question-answering interactions for an asker’s information needs. The current study attempted to focus on identity information in a question and investigated how amount and/or type of identity information relate to responses in SQA. Based on the findings regarding different amounts and types of identity information, further studies should be focused to address information seeking behaviors in various sensitive topics (e.g., income, politics, etc).

**References**


TagPad: Supporting Field Interviews and Analysis

Nis Bornoe  
Aalborg University  
nis@cs.aau.dk

Louise Barkhuus  
Stockholm University  
barkhuus@dsv.su.se

Abstract

The area of cyberinfrastructures has looked extensively at research within the natural sciences, however, the social sciences have been largely overlooked in terms of novel data collection and analysis systems. We developed a probe tool, TagPad, to look at the process for social science data collection through interviews and surveys. Our research participants found that TagPad facilitated structuring of interviews but we also found that the setting in which the interview takes place is essential to the success of using this particular tool. We conclude suggesting future designs of social science research tools.

Keywords: cyberinfrastructures, e-science, e-research, social science, qualitative research

Introduction

Scientific fieldwork poses a range of challenges to researchers, both in terms of conducting data collection and analyzing data collected 'in the wild'. These challenges include choosing methods for data collection, methods for analysis and choosing technologies that will best support the method decisions. The decisions needed should be informed by issues such as appropriate method for the research question and available technology. These technologies need to be tightly tailored to the research aim and question in order to be efficient. Recently the field of cyberinfrastructures (or eScience) has addressed the need for more powerful computational tools for data collection and analysis. With a focus on large scale data sets and natural science, however, the field has broadly left out technologies used by social scientists and their often more qualitatively oriented studies (Wouters & Beaulieu, 2006). Although social science fieldwork often make use of simple technologies for example when conducting ethnographic studies, such as audio recorders and notebooks, and simple affinity diagrams for analysis, these scientists also utilize a wide set of computational tools, many which have hardly been addressed within cyberinfrastructures (Kazianas, Sawyer, & Østerlund, 2011).

In this paper we address the complexities of social scientists tasks of collecting data 'in the wild', particularly in terms of interview data and text data collection as well as the treatment of this data for analysis and result-generation. Previous research has provided insights into many aspects of data collection methods for fieldwork research, but few studies have looked closer at novel data collection technologies in their real word use. We developed a fieldwork probe named TagPad, which is a tool for interview and text collection, as well as a tool for analysis support. We let eight researchers and scholars use TagPad for their studies and observed them using it, as well as interviewed them afterwards about their experiences. We found several factors influencing their use of this probe tool in terms of setting, the structure of interviews and analysis, and describe how their interview and data collection practices in return affected the usefulness of TagPad. The useful aspects of TagPad was its simple option for back rolling (that if need be, it could be used as a regular audio recorder, the previous technology the researchers had used) as well as its smooth blending in to some settings. The tool’s weaknesses showed up in settings that were sensitive to new invasive looking technologies and therefore did not fit into naturally. Instead it prohibited natural interaction between researchers and participants. Before presenting TagPad in detail we review previous literature that has studied research processes.

Acknowledgements: We thank the research groups and individual researchers for their time. This research was funded by NSF Grant #0838330. Bornoe, N., & Barkhuus, L. (2013). TagPad: Supporting field interviews and analysis. iConference 2013 Proceedings (pp. 316-325). doi:10.9776/13204

Copyright is held by the authors.
Related Literature

Cyberinfrastructure Research

Cyberinfrastructure, or e-science, describes a range of technologies such as computational support, access to data sources, tools to support access and use, and analysis of this data. The area emerged as computational possibilities increased and was seen as a ‘neutral’ technological platform for large scale research (Kaziunas et al., 2011). However, much current research on cyberinfrastructure has been dominated by the natural sciences (Schroeder & Fry, 2007). While grid computing, shared access to computational models, or very large data sets, do have major applications to some parts of the social sciences, much social science research tackles quite distinct research problems that render these systems inappropriate (David, 2006). The emerging program of e-social science, or ‘e-research’ has developed to address the distinctive problems of social science research (Wouters & Beaulieu, 2006). Technology aimed at social science is an expanding field and several tools have been presented during recent years (Beaulieu & Wouters, 2009). In social science different Computer Assisted/Aided Qualitative Data Analysis software (CAQDAS) tools have been available for a long time. While acknowledging the great diversity within the social sciences, social science research frequently deals with relatively small sets of data, theoretical explorations, and a long term commitment to investigating from interpretivist, or at least non-positivist, traditions (Barjak et al., 2009; Locke, 2011; Lynch, 2000). For these social scientists their work problems and potential use of cyberinfrastructure are distinctly different from the concerns of ‘big science’. Social scientists rather rely on existing technology such as e-mail and portable tools and focus on robustness of tools (Beaulieu & Wouters, 2009). The study of many complex activities - such as learning or language socialization - involves collaboration that is seldom reducible to sharing raw data, but rather a complex process of ‘coming to see’ the world in common (Barkhuus & Brown, 2012; Brown, Lundin, Rost, Lymer, & Holmquist, 2007; Goodwin, 2001). These practices demand distinctive collaborative tools, tools built from an understanding of the intellectual and collaborative processes involved.

Fieldwork Technologies

With new computational possibilities and mobile devices, the option of using data collection technologies in the field emerged. Already in 1999, Guice, Hoffower, & Norvig suggested three types of potential technologies for fieldwork among astrobiologists, which they had studied. Acknowledging that different classes of work practices require different classes of information technology innovation, they divided the tools into project and logistics management tools, mobile communications tools and advanced data collection and manipulation systems (Guice, Hoffower, & Norvig, 1999). Our TagPad in fact ranged across each type by providing easy management of interview and text data material as well as being a specific data collection tool. It even provides wireless uploading of data (mobile communications tools), although this is obviously more on the account of the platform, the iPad, which allows this, than a feature of TagPad itself. In terms of already existing systems several early experimental systems have been presented. Coughlan et al for example created a remote communication system for earth science teachers and students, where data from the field (voice, pictures, video) was streamed to a “mission control environment” (Coughlan et al., 2011). They report pros and cons of this approach, particularly in terms of the obstacles for the researchers inside to get a detailed enough overview of the data. In terms of collaboration around paper writing services such as Google Docs and other cloud based services now provides similar tools as was suggested in 1999 by (Cohen, Cash, & Muller, 1999): Current status of group members (logged in or not), current status of the document in question and (a limited) history of edits in the document.

Preliminary Study of Fieldwork Practices

The work on TagPad builds on a previous study that we conducted, exploring fieldwork practices and collaboration among social scientists (Barkhuus & Brown, 2012). Our study illustrated how the strongest collaborative research groups are of often small scale, long term groups of 3-4 people who continuously work together on several projects, often through the majority of the researchers’ career. Our
study also pointed out how their successful collaborations were often supported by technologies that were flexible and easy to adjust to the task at hand, for example people using ‘in-between technologies’ such as text-chat and video meetings. Although the study was a broader exploration of social science fieldwork and collaboration it led directly to the design and development of TagPad, particularly based on two core observations: Our participants often found it difficult to structure fieldwork interviews in a way so colleagues could replicate them or at least conduct them similarly enough for appropriate analysis. Secondly, they found it difficult to conduct ongoing analysis during or right after actual data collection. We therefore developed a tool to support interviewing and early stage analysis.

After the initial development we presented the tool to different researchers and interviewed them about their current practices based in actual projects (Bormoe, Barkhuus, Brown, & Hall, 2011). We found that the strategies for both the data collection and analysis were only prepared on a very general scale. This applied to all aspects of a study from recruiting participants, deciding on interview locations and interview strategies. Mainly a general study design was decided beforehand, all remaining parts would be dealt with ad-hoc. Simple tools were used for both collecting and managing the data. For example, word processing software was often used both during the data collection and the analysis. The use of dedicated CAQDAS tools was uncommon. Particularly in the small-scale studies commonly available software was used. Also price of the software was a factor.

Based on our examination we identified two centrals elements of the data collection phase. One element is the actual collection when a researcher conducts an interview or observation and saves data such as field notes, audio, video etc. We believe that this phase can be supported with technology, especially, relative cheap tools not requiring a high level of technical expertise. Another element of the data collection phase is the processing and management of the data. This includes organizing, storing, archiving, and preparing the data for analysis. This part of the data collection was found to be challenging and time consuming. By automating some of these tasks, such as a simplifying data uploading, a goal with introducing a tool is to make this part less challenging.

![Figure 1. Left: An interview guide is loaded into TagPad. Here is question 1.2 selected. Right: In the analysis view tags can be added to each interview for pre-analysis.](image)

**The Design of TagPad**

TagPad is an iPad app designed for researchers conducting interview studies. The application can be used for qualitative studies such as open-ended questions or quantitative data collections such as in-person surveys and is designed both for in-person and phone interviews. It's been designed to fit as flexibly as possible with structured and semi-structured interviews. Our aim was to design a “turnkey” multipurpose research tool that easily can be integrated into a diverse set of studies. A design goal was to
make an “off-the-shelf” app centered around simplicity. Further it was an ambition to develop technology that does not require a high level of technical expertise and easily can be implemented in low budget projects. The goal of making a tool that fits as many studies as possible is also a disadvantage because interview studies exists in all shapes and sizes and often studies will have individual requirements. The app is specifically intended to support the data collection phase and also offers the possibility to add tags to the interviews for pre-analysis. With TagPad the idea is to provide a tool that can support research and automat some processes while still offering the researcher flexibility and creativity so it’s not the tool dictating the data collection.

Conducting a study is time consuming and complex, and often involves researchers with different backgrounds and skills. We attempted to design TagPad with a limited set of features and options, we also made sure it produces output files in common formats. By automating processes such as a one click upload feature we removed complex tasks. We used a common feature could service for data storage. TagPad can also be used with applications using voice over Internet Protocol (VoIP) services such as Skype making it possible to use TagPad for telephone interviews.

Conceptually TagPad can be divided into three different parts: the TagPad app itself, the platform it’s running on – a tablet computer, specifically an iPad, and the integration with a cloud storage service, in this case the service Dropbox.

**The TagPad Application**

The TagPad application has two different views: the interview view and the analysis view, the interview view being the main view (See Figure 1.). As input data TagPad can record audio, save short text entries and use multiple-choice items. The researcher decides how to combine the input. For example, TagPad can be used to record audio only, it can be used for only short text entries and/or multiple-choice questions (or a combination) or a combination of text and audio. Because of this flexibility TagPad can be considered a multi purpose interview tool suitable for a wide selection of qualitative and quantitative studies. TagPad has a simple analysis view for quick analysis to support selective transcription (See Figure 1.). Besides having playback capabilities tags are automatically added to the audio timeline for each question so it is possible to locate where in the audio a specific question was asked. Custom tags can be added making it possible to navigate, locate and share specific audio segments. Tags can be added both during and after an interview.

The portability of the iPad addresses the vital aspect: mobility. Additionally mobility is supported because the iPad agilely can be prepared for an interview and operates fairly long on battery without the need for an external power source. Mobility is essential because data collections in the form of interviews often will take place in the field requiring flexibility of the researcher. The use of cloud storage is also a defining feature. Advantages are effective and simple data maintaining, distribution, and support for collaboration. With cloud storage the service provider handles most of the technical aspects such as backup of data and there is no need for manual management of the storage and configuration of the system. Disadvantages include potential issues with bandwidth, performance, reliability and availability and that some projects may have very specific records-keeping requirements.

**Studying TagPad in Use**

To gain further insights into actual use of the tool we contacted research groups, both local, national, and international and asked if they were interested in using TagPad and if they had any research projects that would be appropriate for its use. TagPad was also made available for free on Apple’s app store. Out of the 50+ groups and individuals we contacted, we found eight researchers or research groups interested in using TagPad (See Table 1.). All studies were either in the study design or data collection phase. We interviewed one representative for each study inquiring into the study outline, time frame and which participants they were looking for. We asked how they would organize the data collection, which methods they would use, by what means they would collect the data and which questions they would ask. We also did document sampling where we looked into the interview guides and
instruments used. We looked into both small and large-scale studies, from research teams consisting of only one person conducting 11 – 20 interviews to research teams of 15+ planning to conduct up to 1000+ interviews. Out of the eight studies, six (LCo, EB, AS, RT, LCa, MO) can be classified as small-scale studies with one to four team members completing 10 – 42 interviews. Two studies (IS, JS) are considered large-scale studies with five+ team members and 100+ interviews already completed or planned.

The studies conducted qualitative interviews, quantitative surveys and observations. A mixed approach was common, such as study observations being mixed with interviews. For two of the studies we were also able to observe the use of TagPad in-situ (LCo, JS). For example we spent a full day observing the field site of the JS study, which provided us with valuable insights into their data inquiry constraints. So although these two researchers only conducted one interview with TagPad before we interviewed and observed them, the information gathered here was very useful.

Table 1
Participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Position</th>
<th>Study details</th>
<th>No. of interviews with TagPad</th>
<th>Analysis with TagPad</th>
<th>Special circumstances</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS</td>
<td>Tenured professor</td>
<td>Qualitative, ~500 participants</td>
<td>5</td>
<td>Preliminar</td>
<td>Children 9 - 10 years old</td>
</tr>
<tr>
<td>LCa</td>
<td>Researcher</td>
<td>Qualitative, ~30 participants</td>
<td>23</td>
<td>Preliminar</td>
<td>Medical field setting</td>
</tr>
<tr>
<td>AS</td>
<td>PhD student</td>
<td>Quantitative survey, 19</td>
<td>19</td>
<td>None</td>
<td>Bars and restaurant</td>
</tr>
<tr>
<td>WL</td>
<td>Master’s student</td>
<td>Qualitative, 20 participants</td>
<td>7</td>
<td>Preliminar</td>
<td></td>
</tr>
<tr>
<td>JS</td>
<td>PhD student</td>
<td>Quantitative /Qualitative, 1000-1500 participants</td>
<td>1</td>
<td>None</td>
<td>Sensitive setting, rarely will participants accept audio recording</td>
</tr>
<tr>
<td>LCo</td>
<td>Post-doc</td>
<td>Qualitative, 11 participants</td>
<td>1</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>EB</td>
<td>PhD student</td>
<td>Qualitative, 12 participants</td>
<td>4</td>
<td>Preliminar</td>
<td></td>
</tr>
<tr>
<td>MO</td>
<td>PhD student</td>
<td>Qualitative, 42 participants</td>
<td>2</td>
<td>None</td>
<td></td>
</tr>
</tbody>
</table>

Analysis

We analyzed our interview data through categorization and constant comparison. With mixed data such as audio-recorded interviews, observational notes and meeting notes, a structured analytical comparison was only possible with a subset of the data.

True to our own approach we actually used TagPad to record our own interviews and utilized the question structure to compare answers across participants. We found this analysis method incredible useful ourselves. Our analysis followed traditional qualitative analysis methods in terms of extracting results from rich text data and categorizing these according to relevant issues.

Our goal was not just to evaluate the usefulness of the TagPad application but also to explore the fieldwork practices through the lens of this potentially influential tool. To avoid any confusion of terminology, we use the terms “the researchers” about our participants and “the participants” about their participants in the studies they conducted.

320
Fieldwork Practices

We first describe how the participants generally had conducted fieldwork prior to being introduced to TagPad and how they would usually record their data. All participants had conducted interviews or face-to-face surveys before, some were also familiar with observational studies. Most had used digital recorders, few had used actual tape recorders and two had used electronic pens which worked as an audio recorder but which also recorded notes taken with the pen, time stamped to the audio. Apart from occasional technical problems the participants were satisfied with their chosen technology. Some had people transcribing the audio for them, others did this themselves and for the surveys, our researchers transferred paper notes to electronic means such as spreadsheets and statistical software for analysis.

Our first experience was that intervening into existing research routines was difficult. As seen in Table 1, only a relative small number of interviews were conducted using TagPad. Several reasons exist: First off in some studies TagPad would be introduced after the beginning of the data collection (IS, WL, JS, MO). Secondly we would conduct interviews with the researchers before the data collection had ended (IS, LCa, JS, MO). Further some researchers found it challenging to implement new technology during a research process. Learning a new tool during the process while also adjusting current routines was a main reason for only using TagPad for a small number of interviews (WL, JS, MO).

The Field of the Field Studies

The participants varied in terms of where they collected their date, some being more focused on ‘the wild’ than others. Three of the studies we looked at were simple interview studies conducted in fairly regulated places such as offices, cafes or outdoors. Other researchers had to seek out their participants in the relevant environment. One study of illegal immigrants for example, was conducted in homeless shelters just across the Mexican border where the target participants were staying. Another study took place in the hospital wards because observation of clinicians’ behavior was part of the data collection and the main interview was conducted right after a ward-round in relation to what the researcher had just observed. To all participants TagPad was their first integrated recording and analysis tool they encountered.

Experiences with TagPad

We now continue describing our participants’ experiences with TagPad. Our focus is not on the usability of the system but instead on what the use of TagPad contributed with or how it affected the studies as well as how our participants were able to adapt it to their specific needs and use, or how they found this difficult. The aim with this description is to be able to conclude on a broader level what type of technologies can be used for interviews and more general data gathering.

Affect on Interview Structure

A defining feature of TagPad is that the audio file and hence the interview is cut into bits according to the interview schedule’s questions. Seemingly flowing and very open-ended interview guides will therefore be forced into at least some sort of structure and we expected researchers who used this approach to perhaps find it constraining to structure their interview more tightly. When interviewing them, some of them in fact expressed such worries but all but one had overcome them. IS for example felt that it was a good way for her to actually stay with the interview guide and felt the interviews became better. She says:

"Yeah, if anything I was worried that (TagPad) would force me to be too structured. But, especially with children. But I'm fine, we go around, I jump from question to question, because I have these chunks, I categorize these chunks with big headings, I mean with what you would do with any interview. (...) No problem (...) and it is actually making it much easier."
A researcher who had already conducted a number of interviews with a regular audio recorder before adopting TagPad, (WL) on the other hand gave up on advancing the questions through selection with her finger during the interviews, and simply did not use the audio bits but instead transcribed the master audio file that TagPad also saves. She reported that she had her interview questions memorized and therefore rarely even looked down at the questions.

More broadly, TagPad worked well with semi-structured interviews, with emphasis on structure. We found that it was essential to have a system that was flexible in terms of being able to ‘roll back’ to previous work practices. With TagPad it was possible to simply use it as an audio recorder, all the participants’ previous tool, in case the forced structure did not fit the data collection situation. But in addition it was an improvement for some of the researchers who found the enforced structure helpful.

Setting Issues

The nature of fieldwork means that data collection takes place in a diverse set of environments and our participants collected data in a multitude of different settings with different effects. Only three of our researchers (LCo, EB, MO) were able to use office settings or settings with a proper table and chair (such as outdoor cafe seating or a common area) for all their interviews. They expressed that there were very few differences in regards to the setting from their earlier interviews, for fairly obvious reasons. The main distinction was that EB had been using a digital pen before and took notes during the interview but now used the iPad; after initial getting used to the new input method, she actually did not feel there was any significant difference regarding the interview setting.

Two researchers (WL, AS) collected data in semi-controlled environments: cafes, restaurants, the beach and other public venues. Here they would often sit on the ground, stand up or even walk around a bit with their participants, increasing the challenges for audio collection and note taking. One factor for a researcher was the monetary value of the iPad; she was not excited about using it on the beach where it might be dropped and sand get into it. For these two researchers though, the iPad generally had a good format; it acted as a notepad with a natural purpose in the setting.

Three of the researchers worked in complex settings that all affected the use and their abilities to use TagPad. One researcher for example collected data at a hospital on a 13 hospital ward, following the medical staff around and recording their activities. She found that TagPad worked really well because people were using clipboards already and she was just yet another person in the room with a notepad. LCa explains how she blended in:

“Because wards are a very busy environment and very challenging to observe in. It was a tool to make me focus well. (...) (E)veryone is carrying clipboards and writing and all look down. And to have something the size of an iPad made people not think to much about it, it made people ignore me because I was doing like everybody else. They weren’t really thinking of me as threatening. It allowed me to blend in in more ways that one.”

This particular setting strengthened the legitimacy of the presence of an iPad and it was not until people noticed it was an electronic notepad and not a paper notepad that they took note and commented on it. It enabled more natural interactions according to our researcher. This example contrasts another study that we followed and in where they used TagPad. This was a large cross-university research study of illegal Mexican immigrants and their experiences with deportation. The method of recruitment meant that researchers spent significant time (several full days per week) at homeless shelters just across the Mexican border and interviewed deported people there. Their previous data collection instrument was paper surveys, which the interviewer filled in, but because this was 30 pages long (with many conditioned questions). They were keen to adapt and transfer the survey to TagPad.
The authors of this paper assisted in transferring the instrument in order to help them along and they acquired several iPads for the study. We followed three researchers to the shelters for one day to observe and assist if any problems arose with TagPad. However, the setting of the study was too complex and what might seem natural in a rich country (using state-of-the-art technology) can be out of place in poor settings with frustrated people. The majority of the participants recruited that day refused to be audio recorded, which is understandable considering their situation but more importantly the researchers reported, and our observer observed, the presence of the iPad was so out of place and seemed intimidating to the participants that they had to take it away. In the end one interview was conducted using the iPad and our interview was with the researcher who conducted this interviews. What is interesting in this situation is that the researchers themselves had not predicted this problem of the inquiry environment being so sensitive to the presence of an iPad. The researchers had all taken the trip several times and were 1/4 through their data collection, yet it was not until the actual situation that they realized the problems and the obstacles that the technology also presented in such sensitive environment.

Participants

The most important part of the fieldwork setting is of course the participants themselves and as we exemplified with the participants from the study of illegal immigrants, their sense of security and comfort means everything for a good interview and data collection. In one study (WL) the researcher reported that TagPad resulted in slightly less eye contact than for example a traditional audio recorder, which was imperative to the nature of her interviews that addressed a very sensitive topic. She explains the diminished eye contact:

“I would prefer just to have a conversation. (...) (B)ecause I am focused on (TagPad) as opposed to a piece of paper where you are just glancing at it, you know, making mental notes in the head because you’re touching this, I feel (...) it takes away the attention for sure, and I don’t like that ‘cause I just want it to (...) have everything in my head.”

For other researchers the eye contact was not a problem because of the training they had as interviewers. It did result in missed advances of the questions with the resulting non-clipping in the audio file but for the ones where this happened, they did not view this as a problem because this simply mirrored their previous practices. One researcher (IS) explained how she was able to effectively integrate TagPad into her interview sessions:

“(...) it took me one interview of having it sitting there like a microphone before I realized I don’t have to do that. It can be less intrusive, it can be part of me and not part of the table.”

Collaboration

Only four of our eight researchers were collaborating with others on their study; they had obviously all collaborated around research before, but for three of them this particularly study was conducted individually, either as part of their individually funded project or as a Masters project. For the studies where more than one conducted the interviews, the forced structure showed useful for sharing data and analyzing results. However, this issue of fieldwork was not looked at in detail in our study of TagPad.

Early Coarse Level Analysis

One of the core features of TagPad was its option for early coarse level analysis, which was assisted by the interviewer creating tags, either ahead of time or on the fly, and adding these to the interviews at specific points in the audio file. When viewing the interview in the TagPad analysis view, it was possible to scan for tags and listen to the bits that were relevant for a particular tag. Although only a subset of the researchers has tot to the stage of analyzing their data with TagPad, most had the option of
this early stage analysis or ‘looking’ through their data. One researcher (LCa) found it very useful to use the tagging to further develop the interview guide and adjust the next interviews. If a couple of interviews had received a particularly high number of specific tags, she knew that this was an area that was important for her study and she could push for that topic in subsequent interviews. As TagPad was not developed for the purpose of extensive analysis we did not expect our researchers to have conducted much analysis that was distinct to their use of TagPad, instead we hoped that they would make use of the separation of audio later on when searching for specific answers to specific questions. However, at the stage of our inquiry, none of our researchers had reached this part of their study.

TagPad as a ‘new technology’

A feature of all new technology and new approaches to established practices is a time of transition where both technologies are often used. Some of us remember the times when record players were lined up next to CD players and many early CD players came with built-in cassette players (the comparison might be unfair because of the cassette tape’s ability to record music in contrast to a CD player, but the example should be illustrative enough). TagPad was to all our researchers a new technology and to some, the iPad itself with its touch screen and gesture based interaction, was also new. This meant that only few had the courage to do interviews solely with TagPad the first times they used it. The common approach was to use the old recording method (tape recorder, electronic audio recorder, note pad) concurrently with TagPad as backup in case of potential technical failure. The main issue is that data collection settings are highly sensitive situations: the researcher has spent a lot of time recruiting and getting the participant and in some instances the participants are taken from a limited pool of potential participants. Missing out a data point can prove crucial to the success of the study and can in most cases result in extra time spent on the project.

Implications for Fieldwork Interview Technologies

Our goal with developing and deploying TagPad has been to provide social science with an assisting tool for qualitative fieldwork, particularly interviewing and surveys. We studied the use of TagPad to gain insights into structures that might affect adoption of such tool and to inform us of further design guidelines of research technologies. We highlight three issues that were apparent after studying TagPad among social science researchers:

1. We found that with new technologies, flexible back-rolling is important in the case that the new method or technology does not work as intended. In our study it helped both researchers who felt the interviews became to structured for their setting and at the cost of intimacy with their participants, and it helped when the technology simply did not work as they had expected. The point is not new in terms of technology adoption, many other technologies rely on easy ‘roll back’ function such as software that can be set to a previous edition for opening older documents, but this is an important point to make in a social science fieldwork setting. For our researchers, participants were ‘expensive’ to recruit in time and effort and any lost data resulted in major frustrations. They did not have an infinitely pool of participants to take from and it was necessary to have options that might not be state-of-the-art inquiry systems, but at least reliable data collection methods.

2. Secondly we found that each research setting is unique and unpredictable. Trying to use a new inquiry technology is always an experiment and it was apparent that in some cases it did not fit into the setting. Instead the researchers had to find other possibilities for data inquiry and in some cases roll back to previous methods. It is not surprising that some settings work with a new tool where others do not, but we pointed to a set of unpredictable factors such as the perceived alienation that the tool contributed to and a sudden sense of value of the system, which resulted in a higher level of sensitivity towards using it in the setting such as the beach where it might get sand into. The tool also promoted eye contact in some situations but almost prevented it in other setting depending on the researcher’s previous experience and perception of the situation.

3. Finally, the researchers benefited from early stage analysis, which they had not been able to do before without listening to the full interviews and writing notes. A tool that can contribute to more detailed and deeper analysis (taking for granted that the early stage results in better and more thorough later analysis) is highly beneficial to social scientists of the type we interviewed and will likely assist in further result generation in the future.
Conclusion

With TagPad, we have developed a potentially beneficial and useful cyberinfrastructure tool for social scientists, an area that has not been addressed as much in as the natural sciences. Our evaluation indicated not just general usefulness but also illustrated how users were able to adjust the tool to their specific study. We believe an app such as TagPad can help streamline and organize the data collection process, making it both easier and less complex to collect data and distribute it. We aim to continue our development and studying of TagPad as well as other research technologies in order to fine-tune more development guidelines for such hopefully useful tools.

References


A Study on Q&A Services Between Community-based Question Answering and Collaborative Digital Reference in Two Languages

Dan Wu  
School of Information Management  
Wuhan University  
woodan@whu.edu.cn

Daqing He  
School of Information Sciences  
University of Pittsburgh  
dah44@pitt.edu

Abstract

To explore further improvement of libraries’ reference services in the Web 2.0 environment, we systematically compared community-based question answering (cQA) sites with collaborative digital reference (cDR) services in both English and Chinese languages. We employed a sampling method where we asked a set of questions of four different types and in three domains at selected cQA and cDR sites. The focus of the study includes evaluation of the answer quality and the responsiveness of the sites to the questions. Our results show that cQA sites provide more answers within shorter response times, and they are probably better suited to answer questions about everyday life or questions with easy answers. In contrast cDR services are augmented by better trained librarians, well organized working procedure, and more extensive information sources, so they produce more effective services. We finally hypothesize ideas of combining cQA and cDR under the goal of further improving cDR services.

Keywords: community-based question answering sites, collaborative digital reference services, social web

Introduction

Community-based question & answering (cQA) sites achieve knowledge sharing among community users through a participatory platform where users can ask and answer questions for each other (Shah, et al., 2009). Since the first cQA site appeared in Korea in 2002, these services have been developing at a very fast pace in many parts of the world. For example, Yahoo! Answers, as by far the most widely used cQA site, had a reported 62 million unique visitors per month in the United States alone in 2010 (Gazan, 2011). As iResearch’s latest survey results show (iResearch), the total number of user accesses on Chinese cQA platforms in April 2010 was more than 2.26 billion times, and it is 1.4 times larger than that of the same period of the year before. All these demonstrate that cQA services have expanded to meet a wide range of people’s information needs, and more importantly, cQA services have been viewed as rapidly developing social collaboration platforms (Shah, et al., 2009).

Collaborative digital reference (cDR), as stated on the American Library Association’s website¹, “extends a library’s information service capability through interaction with other libraries or information centers”. Based on formal established protocols, different libraries and information institutions in cDR work together to provide assistance (online reference and remote document delivery) to users to meet their information needs. Reference librarians and users in those services may come from the same library or sister libraries. As with other forms of online services, the reference services are often conducted via both form-based and live chat-based online reference. With the support of cDR, reference librarians with

¹ http://www.ala.org/rlsa/contact/rosters/rss/rus-moucoop

Acknowledgements: This work was supported by Wuhan International Science and Technology Cooperation Fund under agreement 201070934337, by the 3rd Special Award of the China Postdoctoral Science Foundation under agreement 201003497, and National Science Foundation of USA under the agreement NSF/IIS 1052773. The authors want to thank the student helpers and the anonymous reviewers.


Copyright is held by the authors.
different areas of expertise and working schedules can learn from each other and help each other to better allocate resources according to users’ needs (Wang, 2007).

Community-based Q&A and digital reference (which includes cDR as a special case) are two instances of social question and answering because they both enable people to collaborate in answering questions (Shah, et al., 2009). It is important and interesting to study the differences and connections between these two services. More importantly, as one of cDR’s major challenges is to expand the services to wider ranges of users, it is beneficial to explore the integration of some cQA’s ideas and functions into cDR to enable cDR to serve even more diverse groups of web users.

In this paper, therefore, we will study a set of cQA sites and cDR services to examine their differences and similarities. In addition, because of our backgrounds, we are interested in exploring the differences between Chinese sites and those services based on English. To achieve these research objectives, we designed our study as a comparative analysis of several representative cQA sites and cDR platforms, and deliberately selected both cQA and cDR sites in Chinese as well as in English. We aim to learn more from the effective practices in cQA sites, and to improve libraries’ reference services in Web 2.0 environment.

In the remainder of this paper, we will first review the literature, then discuss the research design of our study, in which a set of selected cQA and cDR sites will be briefly presented. Next, we will describe the results obtained from the studies, discuss obtained insights about the differences and the connections between cQA and cDR services, and explore ideas for applying cQA functions in cDR. Finally we will end with our conclusions and thoughts on future work.

Related Work

Community-based question and answer (cQA) is also known as social QA in the literature (Gazan, 2011). Ever since Knowledge-iN launched in 2002 in South Korea as the first cQA, the study of cQA has been an active research topic, which can be classified into two categories (Shah, et al., 2009). The first research area focuses on the user generated content, such as questions and answers, and pays particular attention to the quality of the answers. For example, by analyzing 81 questions they posed in Yahoo! Answers, Su et al. (2007) demonstrated that the answer quality in the cQA site varied widely. Harper et al. (2008) compared the quality of answers in cQA to those of other types of online Q&A services, such as digital reference services and expert services, and found that cQA answers are perceived as being of higher quality. Liu et al. (2008) examined answers in cQA, found that user satisfaction is an important quality factor, and thus developed several algorithms to predict user satisfaction with the answers.

The second area of research is on users in cQA, which include questioners, answerers and the cQA community as a whole. But majority of the studies paid attention only to questioners and answerers. Gazan (2006) classified the answerers as being either specialists in a particular topic or synthesists who can combine several schools of thoughts and knowledge into one coherent presentation. He found that questioners prefer both types of answerers in order to balance between expertise and information in received answers. Interestingly, studies also showed that answerers select questioners as well, and prefer to respond to the questioners who actively seek for information (noted as seekers) rather than those who just want some ready text for completing tasks such as class assignments (denoted as sloths) (Gazan, 2007). There are many works on identifying authoritative users in the community (Bouguessa, Dumoulin, & Wang, 2008; Jurczyk & Agichtein, 2007).

Digital reference extends traditional library reference services into the Internet environment, and it has resulted in a great deal of discussions (Janes, 2002; Lankes, 2004; Pomerantz, Nicholson, Belanger, & Lankes, 2004). The central question of digital reference, as Lankes (2004) states, is “how can human expertise be incorporated effectively and efficiently into information systems to answer information seekers’ questions?” In their report about building syncronicity for digital reference, Connaway and Radford (2011) state that the challenges for the further development of digital reference is that users still do not really know about digital reference services, and users satisfaction plays critical role in the success of a positive digital reference experience. They identified that the two key factors for users’ satisfaction are the information provided in the answers and the interpersonal relationships between librarians and the users. They recommended that libraries should cooperate more with peer institutions and other services.

Community-based Q&A and digital reference services can be viewed as two instances of online Q&A (Shah, et al., 2009). Therefore, researchers identified the shared common factors, as well as their
differences. For example, Shachaf (2009b) notes that while librarians are generally regarded as valuable sources for answering more difficult queries, cQA sites are more heavily utilized. Researchers are also interested in the combination of the two services (Shachaf, 2010). Lankes (2004) envisioned a "general digital reference model" that contains a Q&A archive. It could be implemented through the cQA service.

The aforementioned studies all concentrated on English cQA sites and cDR services. There is also much research on Chinese cQA and cDR services (Wang, 2007; Zhang & Yuan, 2009), however, no existing study systematically compares English and Chinese cQA and cDR services.

Research Design

Research Questions

Based on the research goal and objectives identified in this study, we propose the following research questions to be answered through experimentation:

- Q1: through the set of questions asked at the selected cQA and cDR sites, what service differences can be obtained in terms of answer quality, answer responsiveness and response time?
- Q2: Do Chinese sites and English sites reveal differences in the answers to Q1?
- Q3. What can be learned from cQA to improve cDR?

Sites Selected for Our Study

As stated, we selected several representative cQA and cDR services in both Chinese and English. They represent the state of the art of the services we could utilize. These services can be classified into the following four types: three Chinese cQA sites (Baidu Zhidao, Sina iAsk and SOSO Ask); three English cQA sites (Yahoo! Answers, Answers.com and MadSci Net); three Chinese cDR sites (the Collaborative Reference Service of China’s National Science Digital Library, Online Collaborative Knowledge Navigation and the Collaborative Reference Network of Zhongshan Library at Guangdong Province); and three English cDR platforms (QuestionPoint, IPL2 and Ask a Librarian). Below, we briefly introduce these sites.

**Baidu Zhidao** is an interactive online question and answering platform. Launched in June 2005, it is the most popular Chinese cQA site measured by both usage volume and service maturity. **Sina iAsk** is a subsidiary online Q&A service of Sina.com, which is one of the most popular online portals in China. **SOSO Ask** is an interactive Q&A platform developed by Tencent.com. As an integrated service of QQ, the most popular online communication/social network service in China, users of SOSO Ask can login with their QQ accounts.

**Yahoo! Answers** is the most popular English cQA site with an active user community as well as well-regarded reputation. **Answers.com** is another very popular interactive Q&A community which claims to have over 60 million user generated questions across 8,000 categories. **MadSci Net** offers over 36,000 answered questions in the domain of natural sciences. It is popular because their questions are answered by scientists, educators and engineering from all over the world.

**The Collaborative Reference Service of China’s National Science Digital Library** is a well known service provided by the National Science Library of China. Users can obtain reference services through three methods: live chat, completing a reference request form and phone-based reference. **Online Joint Knowledge Navigation** is a cooperative service led by Shanghai City Public Library (another famous library in China), and includes public libraries, academic libraries, and research libraries in the Shanghai region as well as libraries from other cities in China. **The Collaborative Reference Network of Zhongshan Province**

---

2 BaiduZhidao: [http://zhidao.baidu.com](http://zhidao.baidu.com)
4 SOSO Ask: [http://wenwen.soso.com](http://wenwen.soso.com)
5 Yahoo! Answers: [http://answers.yahoo.com](http://answers.yahoo.com)
6 Answers.com: [http://www.answers.com](http://www.answers.com)
7 MadSci Net: [http://www.madsci.org](http://www.madsci.org)
8 The Collaborative Reference Service of China’s National Science Digital Library: [http://www.csdl.ac.cn](http://www.csdl.ac.cn)
9 Online Joint Knowledge Navigation: [http://zsdh.library.sh.cn:8080](http://zsdh.library.sh.cn:8080)
Library at Guangdong Province is a library cooperative consortium led by Zhongshan City Library in Guangdong Province and includes libraries from all over China. It has one of the largest Chinese digital resources, and offers free online reference services and remote delivery service to people in China.

QuestionPoint is a well known global collaborative digital reference platform that has a very large group of participants and offers a wide range of services. It is managed by OCLC and the Library of Congress of United States. As a collaborative service, users of QuestionPoint can direct a question to their local library, and the answers are provided by the librarians with expertise in the relevant subject areas, regardless of their actual location. In our study, we chose the “Ask a Librarian” service offered by the Library of Congress. IPL2 is currently maintained by the School of Information Science and Technology at Drexel University, and it is based on a well-received online digital reference website developed by Michigan University in 1995. IPL2 combines IPL (Internet Public Library) and LII (Librarians' Internet Index). Ask a Librarian is a virtual reference system developed through the cooperation of the local university libraries and public libraries in Florida. Over one hundred librarians work together to offer this service to their users.

Table 1: Four types of questions on three different domains

<table>
<thead>
<tr>
<th>Factual questions</th>
<th>Enumeration questions</th>
<th>Definition questions</th>
<th>Explorative questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>芒德尔-托宾效应最早是在哪篇文章中被提出？</td>
<td>在所有诺贝尔文学奖得主中，有哪些人是从南美洲来的？</td>
<td>什么是流动性补偿？</td>
<td>全球经济复苏还需要多长时间？为什么？</td>
</tr>
<tr>
<td>In which paper was the idea later called Mundell-Tobin effect first published?</td>
<td>Among all the Nobel Literature Prize laureates, who were from South America?</td>
<td>What does compensation for liquidity mean?</td>
<td>How much time is still needed for global economy to recover? Why?</td>
</tr>
<tr>
<td>迄今为止，诺贝尔文学奖已有多少位获奖者？</td>
<td>世界图书首都评选是从哪一年开始的？</td>
<td>什么是泛文学?</td>
<td>博客对大众文学有哪些影响？</td>
</tr>
<tr>
<td>How many people have won the Nobel Prize for Literature up to now?</td>
<td>From which year did the selection of &quot;World Book Capital&quot; begin?</td>
<td>What does pan-literature mean?</td>
<td>What impacts have the blogs made on the popular literature?</td>
</tr>
<tr>
<td>世界性的图书馆组织有哪些?</td>
<td>世界的图书馆组织有哪些?</td>
<td>什么是iSchool?</td>
<td>数字图书馆的快速发展会给实体图书馆带来哪些方面的重大变化？为什么会有一些变化？</td>
</tr>
<tr>
<td>What international library organizations are there?</td>
<td>What is iSchool?</td>
<td>What is iSchool?</td>
<td>What important changes will the rapidly developed digital libraries bring to traditional libraries? And why there are these changes?</td>
</tr>
</tbody>
</table>

Study Design

The research presented consists of a study based on sampling method, whose purpose is to obtain first-hand, focused evaluation of the selected sites. To guarantee adequate coverage for the

---

10 The Collaborative Reference Network of Zhongshan Library at Guangdong Province: [http://www.ucdrs.net](http://www.ucdrs.net)
11 QuestionPoint: [http://www.questionpoint.org](http://www.questionpoint.org)
12 IPL2: [http://www.ipl.org](http://www.ipl.org)
13 Ask a Librarian: [http://www.askalibrarian.org](http://www.askalibrarian.org)
study, we selected four types of commonly asked questions: factual questions whose answers are about certain specific items, enumerative questions whose answers should contain all the related points, definition questions whose answers contain a definition of certain concept and explorative questions which are related to some complex issues that often lack definite answers. At the same time, we chose three different domains that are relatively diverse: economics, literature and library science. Therefore, in total, we issued 12 questions to each cQA or cDR site (see Table 1). Considering that there are both English and Chinese sites, all questions are submitted either in English or in Chinese accordingly. When conducting experiments on those Q&A sites, we asked the questions one-by-one according to the regulations of those sites, whereas during interactions with cDR sites, we tested the request form reference service because of its closer resemblance to cQA than the live chat reference service. Since MadSci Net only accepts questions related to natural sciences, we had to redesign the questions specifically for this system: 1) What is the fastest bird in the world? 2) How many types of conventional clean energy are there? 3) What is SOFC? and 4) Why does America always suffer from hurricane? When the correctness and quality of the obtained answers were judged, the authors and their student helpers looked at printed encyclopedias, Wikipedia and online fact books for answers to factual, enumerative and definition questions, and consulted domain experts on answers to explorative questions. This study started on July 14 2010 and finished on August 25 2010: thus, it lasted for 43 days.

Results Analysis
Chinese Experiment Results Analysis
Table 2 shows the number of correct answers and the total number of returned answers for each of the questions. In total, we received 43 answers for the 12 questions asked at the cQA sites; on average 3.58 answers per question. This average number is much smaller than that for the most popular questions on these cQA sites. The reason might be because our questions are more or less related to academic topics. However, that number is still higher than that of answers received from the cDR sites (a total of 29 answers for the same 12 questions with an average of 2.42 answers per question). This means that the cQA sites provided on average one more answer for each of the 12 questions than the cDR sites. In terms of the number of correct answers, cQA produced 33 out of 43 (76.7%), whereas cDR had 23 out of 29 (79.3%). So in summary, the cQA sites were more effective than cDR in eliciting the number of answers, but cDR sites had a higher level of accuracy in their answers.

When evaluating the responsiveness and answer quality based on question types, factual questions generated the lowest number of answers (5 answers for 9 questions) from the cQA sites. This might be due to some nature of the questions themselves. For example, the economics question received no answers from the three sites at all. Our factual questions might be difficult in finding answers for the users in these sites: only 4 questions were correctly answered. In contrast, the cDR sites produced 8 answers for the 9 questions, of which only one is wrong.

The numbers of total returned answers and correct answers to our enumerative questions from the three cQA sites were 13 and 11 (so the accuracy rate is 84.6%), and that from the three cDR sites were 8 and 7 (so the accuracy rate is 87.5%). Similarly, the three cQA sites produced more answers than the three cDR sites on the definition questions (14 vs. 8), and more correct answers (10 vs. 7), but the accuracy rate of the answers was lower (71.43% vs. 75%). These differences are consistent with the literature which notes that the experienced staff at cDR sites have higher information literacy skills, and stronger abilities in searching for information in particular domains.

To our surprise, although the answers for the explorative questions from the cQA sites were similar to those of other types of questions (a total of 11 answers with 8 of them correct), the responses from the cDR services were really low. Only 6 answers were returned, of which 2 were correct. The low accuracy rate of the answers from cDR might be related to our criteria for judging the answers: we count the articles that readers cannot obtain as incorrect answers because common users often cannot access certain databases to download the paper. However, the low response rate from the cDR sites really deserves further study in the future. Maybe it is because the explorative questions we asked are not suitable for these digital references?
Table 2: Number of correct answers/total number of answers to the Chinese questions

<table>
<thead>
<tr>
<th>questions</th>
<th>cQA sites</th>
<th>cDR sites</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baidu</td>
<td>Zhidao</td>
<td>Sina</td>
<td>SOSO Ask</td>
<td>The Collaborative Reference Service of China's National Science Digital Library</td>
<td>Online Joint Knowledge Navigation</td>
<td>The Collaborative Reference Network of Zhongshan Library at Guangdong Province</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual questions</td>
<td>Economics</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
<td>0/1</td>
<td>1/1</td>
<td>0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>0/1</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Library</td>
<td>0/0</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enumerative questions</td>
<td>Economics</td>
<td>1/1</td>
<td>1/2</td>
<td>2/2</td>
<td>1/1</td>
<td>1/1</td>
<td>0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>0/0</td>
<td>1/1</td>
<td>2/2</td>
<td>0/0</td>
<td>1/1</td>
<td>1/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Library</td>
<td>1/2</td>
<td>1/1</td>
<td>2/2</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition questions</td>
<td>Economics</td>
<td>1/1</td>
<td>1/1</td>
<td>2/2</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>1/2</td>
<td>1/2</td>
<td>1/2</td>
<td>0/0</td>
<td>1/1</td>
<td>1/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Library</td>
<td>1/2</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>1/1</td>
<td>0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explorative questions</td>
<td>Economics</td>
<td>0/0</td>
<td>1/1</td>
<td>3/3</td>
<td>1/1</td>
<td>0/1</td>
<td>0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>1/2</td>
<td>0/0</td>
<td>1/2</td>
<td>0/0</td>
<td>1/1</td>
<td>0/1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Library</td>
<td>1/2</td>
<td>0/0</td>
<td>1/1</td>
<td>0/1</td>
<td>0/1</td>
<td>0/0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 summarizes the collected data including the number of answers from each site, the time it took to answer the questions, rate of correct answers, and ranking of the systems based on the correct answer rate.

Table 3 shows that SOSO Ask and Online Joint Knowledge Navigation are ranked as the top two Chinese services. SOSO Ask responded relatively quickly (the shortest response time among the three cQA sites) and produced the highest number of answers. Thus, it provided faster service as well as higher quality service. Online Joint Knowledge Navigation provided answers to all of the 12 questions, even though the answer accuracy rate is not as high as SOSO Ask. Its answering time was very short too. In contrast, although the Collaborative Reference Network of Zhongshan Library at Guangdong Province had the shortest answering time among all three cDR services, these answers suffered from low quality. After examining the answers provided by the Collaborative Reference Network, our impression is that some answers are too simple or coarse, and that the quality depends too much on the training of individual librarians. However, we acknowledge that their service was still better than that of non-professionals. In our study, another surprising finding is that cQA sites were not (as people had expected) faster at providing answers when compared to the cDR sites. The time to answer of the cQA sites to certain questions could be much faster; however, the overall answering time was comparable for both types of services because some questions posted in cQA sites were answered late or even did not receive any answer at all.
Table 3: Evaluation of the Chinese Sites

<table>
<thead>
<tr>
<th>rank</th>
<th>system/Q&amp;A websites</th>
<th>number of questions that received answers (out of 12 questions)</th>
<th>number of correct answers/total number of answers</th>
<th>correct answer rate (%)</th>
<th>answering time (average over all returned answers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SOSO Ask</td>
<td>8</td>
<td>17/19</td>
<td>89.5</td>
<td>1 day, 20 hours and 3 minutes</td>
</tr>
<tr>
<td>2</td>
<td>Online Joint Knowledge Navigation</td>
<td>12</td>
<td>10/12</td>
<td>83.3</td>
<td>3 days</td>
</tr>
<tr>
<td>3</td>
<td>Sina.iAsk</td>
<td>8</td>
<td>9/11</td>
<td>80</td>
<td>13 days, 19 hours and 5 minutes</td>
</tr>
<tr>
<td>4</td>
<td>The Collaborative Reference Service of China's National Science Digital Library</td>
<td>9</td>
<td>7/9</td>
<td>77.7</td>
<td>7 days</td>
</tr>
<tr>
<td>5</td>
<td>The Collaborative Reference Network of Zhongshan Library at Guangdong Province</td>
<td>8</td>
<td>6/8</td>
<td>75</td>
<td>8 hours</td>
</tr>
<tr>
<td>6</td>
<td>Baidu Zhidao</td>
<td>8</td>
<td>7/13</td>
<td>53.8</td>
<td>6 days and 15 hours</td>
</tr>
</tbody>
</table>

Table 4: Number of correct answers/total number of answers to the English questions

<table>
<thead>
<tr>
<th>questions</th>
<th>cQA sites</th>
<th>cDR sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yahoo! Answers</td>
<td>Answers.com</td>
</tr>
<tr>
<td>Factual questions</td>
<td>Economics</td>
<td>0/0</td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>1/1</td>
</tr>
<tr>
<td></td>
<td>Library science</td>
<td>0/0</td>
</tr>
<tr>
<td></td>
<td>Natural Science</td>
<td>0/0</td>
</tr>
<tr>
<td>Enumerative questions</td>
<td>Economics</td>
<td>1/1</td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>Library science</td>
<td>1/1</td>
</tr>
<tr>
<td></td>
<td>Natural Science</td>
<td>0/0</td>
</tr>
<tr>
<td>Definition questions</td>
<td>Economics</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>0/1</td>
</tr>
<tr>
<td></td>
<td>Library science</td>
<td>1/1</td>
</tr>
<tr>
<td></td>
<td>Natural Science</td>
<td>0/0</td>
</tr>
<tr>
<td>Explorative questions</td>
<td>Economics</td>
<td>2/2</td>
</tr>
<tr>
<td></td>
<td>Literature</td>
<td>0/1</td>
</tr>
<tr>
<td></td>
<td>Library science</td>
<td>2/3</td>
</tr>
<tr>
<td></td>
<td>Natural Science</td>
<td>0/0</td>
</tr>
</tbody>
</table>
English Experiment Results Analysis

Table 4 shows the results for the six English systems. The results are organized based on the sites and four question types.

Again to our surprise, two of the five English services did not produce any answers to our questions. Within the three cQA sites, only Yahoo! Answer generated multiple answers to some of our questions (15 answers to 10 of the 12 questions). Answers.com did not return answers to any of our 12 questions. We only asked 4 questions at the MadSci.Net site, but only one was answered, which was not correct. Among the three cDR services, IPL2 dutifully provided one and only one answer to each of our questions, which were all correct. Library of Congress handled all three of our factual questions well, but produced only one answer to each of the other types of questions. Ask a librarian was even worse: it did not return any answers at all.

There was only one answer to the factual questions from the 3 cQA sites, whereas the response rate and correct answer rate of the cDR were both 100% except for Ask a Librarian. This again demonstrates that cDR services have more reliability in providing services.

Yahoo! Answer produced four answers to the three enumerative questions, three of which were correct. The three cDR sites also produced four answers, all of which were correct. The correct answer ratio to the definition questions from the cQA sites was 2 out of 4 whereas that from the cDR sites was 4 out of 4. This shows that the cDR sites had a higher quality of answers.

The correct answer ratio to the explorative questions on the cQA sites was 4 out of 6, and that of the cDR sites was 3 out of 4. Again, the level of responsiveness and the correctness of the answers both depended on the strong performance of a single service, which in the case of cQA was Yahoo! Answers, and that of cDR was IPL2. Specifically, IPL2 answered correctly all the questions we proposed to them.

As with the results for the Chinese systems, we summarize the data from the English systems, and rank the systems based on the correct answer rate (see Table 5).

Table 5: Evaluation of the English Systems

<table>
<thead>
<tr>
<th>rank</th>
<th>systems or sites</th>
<th>number of questions that received answers (out of 12 questions except MadSci.Net which was 4 questions)</th>
<th>number of correct answers/number of answers</th>
<th>correct answer rate (%)</th>
<th>answering time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IPL2</td>
<td>12</td>
<td>12/12</td>
<td>100</td>
<td>14 days</td>
</tr>
<tr>
<td>2</td>
<td>Library of Congress</td>
<td>6</td>
<td>5/6</td>
<td>83.3</td>
<td>17 days</td>
</tr>
<tr>
<td>3</td>
<td>Yahoo! Answers</td>
<td>10</td>
<td>10/15</td>
<td>66.7</td>
<td>2 days</td>
</tr>
<tr>
<td>4</td>
<td>MadSci Net</td>
<td>1</td>
<td>0/1</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>5</td>
<td>Ask a librarian</td>
<td>1</td>
<td>0/0</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>6</td>
<td>Answers.com</td>
<td>0</td>
<td>0/0</td>
<td>0</td>
<td>/</td>
</tr>
</tbody>
</table>

IPL2 is the best online service we have encountered in this study. It not only answered all of our questions with the highest possible correct answer rate (100%), its librarians gave us answers with detailed and accurate information. They even provided resolution steps to the problems and screening methods for examining the collected information. The reference service at the Library of Congress only answered half of the questions with a very long answering time. But it offered remote document delivery service for free, which is rare among the reference services. Among all of the English cQA sites, Yahoo! Answers has the fastest answering speed, and the largest number of answers. However, the quality of its answers is not as good as IPL2 and Library of Congress. Answers.com and Ask a Librarian produced a negligible level of answer or response. One reason for the poor performance in Answers.com might be that it has many users asking questions, so our questions might not be the ones that users would like to answer. However Yahoo! Answers is an even more popular cQA site, and we received answers from that site. Ask a Librarian might primarily serve only local people. MadSci.Net did not give us much service at all.
Comparison between Chinese and English Results

Before the experiments, we hypothesized that there could be service variations between the Chinese sites and the English sites due to their language and cultural differences. Our results demonstrate that this hypothesis is both true and false to some degree.

Although in different languages, the sites we studied exhibit many similarities. For example, using Yahoo! Answers as the representative of English cQA sites to compare with the three Chinese cQA sites, all four of them performed well on enumerative questions, definition questions and to some degree explorative questions, but poorly on factual questions particularly economics. Similarly the cDR sites in both languages were more reliable in providing their services, and produced higher quality answers even though the total number of answers was smaller in comparison to that of the cQA sites. Also, among the four types of questions, the cQA sites from both languages produced slightly better quality answers on factual questions and enumerative questions, and slightly lower quality answers to definition and explorative questions. However, the number of answers to the factual questions was much lower than those to other types of questions. Similarly, the cDR sites in both languages performed almost identically among factual, enumerative and definition questions, but poorly on explorative questions both in terms of answer quality and the number of answers.

However, we do see some differences between the sites in the two languages. First of all, probably due to our authors’ background, our questions to the Chinese sites produced more responses from both Chinese cQA and Chinese cDR sites, whereas two English cQA sites and one English cDR site did not answer our questions at all. In addition, although the English cQA site Yahoo! Answers answered our questions in a timeframe comparable to the shortest response time among the Chinese sites, the two English cDR sites were among the longest response time of all of the tested sites. Their time was at least double of the longest response time of the Chinese cDR sites. Maybe it is due to the fact that both IPL2 and Library of Congress have very busy reference services.

Discussion

Evaluation of Answers’ Quality

Considering the overall experiment results, the cQA sites produced more answers in a shorter time. This is closely related to these sites’ greater numbers of users, and higher level of participation. Among the four types of questions, the cQA sites are better at enumerative questions and definition questions in terms of the number of answers and answer quality. The possible reason for this is that these two types of questions benefit from collective wisdom, and more answers can offer a higher likelihood of getting the correct answers to the questions. However, for questions requiring professional and factual knowledge, users in these sites may not want to spend their time in searching for the answers without a reward, so these types of questions would generate fewer responses (such as in factual questions) or less accurate answers (such as in explorative questions). Therefore, cQA sites are probably better suited for answering questions about everyday life or questions with easy answers. We also noticed that some users in the cQA sites just copied and pasted answers to explorative questions from different sites or composed responses simply based on their own knowledge without conducting more thorough research.

Because often only one librarian in a cDR site is allocated to answer a question, there are limited numbers of answers and less answering time the librarian can manage. Therefore, cDR sites often return lower number of answers and take longer time to deliver the answers. However, with much more authority and extensive information sources, cDR sites can produce higher quality answers. The limitations on cDR also include that some questions were ignored by reference librarians in some cDR sites, which cannot be easily compensated as in cQA sites by a large number of people. We also noticed that reference librarians, probably due to training, like to offer related resources in their answers rather than answering the questions directly. This helps to establish the authenticity and authority of the answers, but it sometimes frustrates the users who cannot easily access the identified resources.

In summary, we think that the advantages of cQA sites include large user groups, highly participative activities, and strong interactions. However, there are also some limitations such as information of different qualities and the shallowness of some answers.
The advantages of cDR systems are that they feature rich and reliable reference resources, and high literacy skills of reference librarians. But they also have limitations such as slow response speed and smaller numbers of answers.

**Inspirations for Libraries from Online cQA Communities**

Through our analyses of these experiment results, we think that the library community can learn from the practices of online cQA communities, and develop better reference services along the following angles.

cQA sites make all of their questions and answers openly available on their websites. Both the questioning and answering processes are open as well. Previous questions and answers are accessible through either browsing or searching. Considering that cDR sites still lack human resources, it might be a reasonable solution to make the cDR reference process and results as open as possible. Of course, we acknowledge the privacy concern from both the users’ and librarians’ point of view. However, some reference questions are identical to those asked in cQA sites, so they can be made public.

Another benefit of making as many reference questions and their answers openly available is to take advantage of the commenting, tagging and discussing functions that are commonly available in social web 2.0 environments. This not only helps to enrich the answers, to encourage deeper post-reference services, but also makes it possible to develop reward mechanisms that have been demonstrated to be very effective in community-based Q&A sites. Of course, we acknowledge the complexity and caution of applying that idea in a library setting.

Our study shows that cQA sites can produce answers of reasonable quality to some types of questions. Considering the time delay in providing cDR services, would it be possible to explore the usages of cQA answers in cDR services? This not only helps to hasten the answering time, but also helps to integrate more closely the cooperation between cQA and cDR services.

Our results show that reference services have great advantages in solving complex professional problems, but they cannot handle the same number of questions as cQA sites received, nor do they provide the same level of users’ participation as in cQA sites. This reflects the idea that cDR is still suffering from a lack of use and awareness in society. Could we learn from successful practices of hosting library-related Facebook pages and YouTube videos, and suggest that some high quality cDR services make them available in well-known cQA sites? This would modernize the image of library reference services and make cDR be recognized by a broader community, particularly groups of younger users. In addition, cDR has the great advantage of experts and information resources which cQA sites cannot compete with, so libraries will have greater potential if they make good use of this advantage. Perhaps cDR may even consider fully integrate the cDR services into cQA sites, so that the strengths of cQA can be combined with the strengths of cDR services.

Although outside of the scope of this study, many cDR sites do provide instant reference services to their users so that the users can get direct and timely online services within a short time limit. This would help to set up cDR sites as different and better service than cQA sites, so it behooves cDR sites to continue to maintain and further explore instant reference services.

We do find that IPL2 is a model cDR service in terms of answer quality and responses to our questions. We hope that it can continue to expand its services to even higher levels, and that other cDR services can learn from IPL2.

**Limitations of Our Study**

We want to make notes on several limitations of our study. First, the number of our study samples is really small considering the popularity of cQA sites and many other cDR services. Although the sites we selected are among the most popular or recognizable, more wide range of sampling would enhance our results and insights. Second, although it helped us to make direct cross-site comparisons, our research method (using a small set of questions we composed) runs the risks of bias with small samples. As we mentioned in result analysis, our selected questions and our native language might trigger or prevent some responses from the English sites. We intend to examine further on these services using naturally occurred questions and answers recorded in their transaction logs. The outcome of that future study would be a good addition. Third, it would be better to have a survey associated with the questions we
asked so that some reasons behind certain reactions from the sites (such as lack of returned answers to our questions) can be better explained.

Conclusions

Community-based Q&A (cQA) and collaborative digital reference (cDR) services can be viewed as two examples of online Q&A; thus, the study of their similarities and differences provides interesting topics for researchers. In this paper, we examined their connections and differences through a sampling research method. We selected a set of famous cQA sites and well-known cDR sites, and utilized them with carefully designed questions that cover four common question types and three domains. In order to fully take advantage of the authors’ background, we studied both Chinese and English cQA/cDR sites. Our results show that the cQA sites produced more answers in a shorter period, which might be due to these sites’ significant numbers of users and their high level of participation. Among the four types of questions, the cQA sites are better at enumerative questions and definition questions, but questions requiring more professional and factual knowledge generated fewer responses (such as factual questions) or less accurate answers (such as explorative questions). Therefore, cQA sites are probably better suited for answering questions about everyday life or questions with simple answers.

Our results with cDR sites show that the number of answers produced by these sites are relatively low and their time delays are often higher in comparison to that of cQA sites. This is compensated for by the expertise of librarians, by the authority and extensive information sources under their control, so that the cDR sites in the end can produce higher quality answers. However, we did notice that some librarians, probably due to training, like to offer related resources in their answers rather than answering the questions directly. This helps to establish the authenticity and authority of the answers but sometimes frustrates the users who cannot easily access the identified resources.

In our discussions, we hypothesize that integrating cQA functions or characteristics into cDR services, will further improve users’ satisfactions with cDR services.

Our future work includes the ideas mentioned in the discussions, such as more extensive exploration of the answering of explorative questions in these two types of services, and careful examination of the reasons for the lack of responses from certain sites. More importantly, we would like to implement some discussed ideas that integrate cQA functions into cDR services.

References


VizDeck: Streamlining Exploratory Visual Analytics of Scientific Data

Daniel Perry
dbperry@uw.edu

Bill Howe
billhowe@cs.washington.edu

Alicia M. F. Key
amfkey@gmail.com

Cecilia Aragon
aragon@uw.edu

University of Washington

Abstract

As research becomes increasingly data-intensive, scientists are relying on visualization very early in the data analysis cycle. We find that existing tools assume a “one-at-a-time” workflow for creating visualizations and impose a steep learning curve that makes it difficult to rapidly create and review visualizations. At the same time, scientists are becoming more cognitively overloaded, spending an increasing proportion of time on data “handling” tasks rather than scientific analysis. In response, we present VizDeck, a web-based visual analytics tool for relational data that automatically recommends a set of appropriate visualizations based on the statistical properties of the data and adopts a card game metaphor to present the results to the user. We describe the design of VizDeck and discuss the results of a usability evaluation comparing VizDeck with three other popular visualization tools. We then discuss design considerations for visualization tools focused on rapid analysis based on observed sensemaking processes.

Keywords: information visualization, visual analytics, user interfaces, scientific collaboration

Introduction

Advances in high-throughput data acquisition technologies have transformed many fields of research from data-poor to data-rich. In turn, the need to analyze potentially large datasets from a variety of different sources has increased the cognitive load of researchers and scientists, taking time away from core research activities. While visual perception has been found to be the highest bandwidth channel from the computer to the human cognitive centers (Ware, 2004), many existing visualization and visual analytics tools require a training period that makes it time-consuming for researchers to bootstrap their usage. Moreover, rapid assessment of new data sources requires data triage: quick assessment of the utility, quality, gross patterns, and internal relationships of the data. This task typically requires rapid review of a large number and variety of related visualizations. Existing tools appear to be ill-equipped for data triage tasks, in that they force the user to spend time designing each visualization created, limiting simultaneous review of several visualizations at once. The question of which types of interactive visual interfaces maximize human efficiency and exploratory analysis is still open, despite a great deal of recent work in this area. As scientists and researchers become ever more cognitively overloaded, the discovery of the best techniques to rapidly leverage human pattern recognition and visual explorative abilities is becoming progressively more crucial.

Research has shown that cognitively overloaded users prefer a simple interface optimized for speed and ease-of-use over a feature-rich interface with a steeper learning curve (Aragon & Hearst, 2005). Additionally, the growth of scientific collaborations that span multiple research groups and disciplines often means that analysis tools already integrated into a researcher’s workflow might not be the same across groups. This effect is measurable: in some scientific collaborations, researchers report that the ratio of time they spend “manipulating data” as opposed to “doing science” is approaching 9 to 1 (Hey, Tansley, & Tolle, 2009). As a result, researchers increasingly rely on visual browsing techniques for exploratory analysis to quickly identify patterns for deeper study.

Acknowledgements: This work was supported by the Gordon and Betty Moore Foundation and NSF award #1064505. Perry, D., Howe, B., Key, Alicia M. F., & Aragon, C. (2013). VizDeck: Streamlining exploratory visual analytics of scientific data. iConference 2013 Proceedings (pp. 338-350). doi:10.9776/13206

Copyright is held by the authors.
This technology gap for rapid exploratory analysis exists in a variety of domains. Astronomers are beginning to build mash-ups to support ad hoc analysis of object catalogs (ASCOT, 2012). In the life sciences, extemporized integration of laboratory assays, public databases, and computational simulation has become commonplace (Howe et al., 2011). Oceanography is transitioning from an expeditionary science to an observatory-based one, where data is streamed real time from an array of sensors (Baptista, Howe, Freire, Maier, & Silva, 2008), decreasing the time available for analysis. In our own experience, when we suggested sophisticated visual analytics tools to support this interaction style (e.g., Tableau), we found that our collaborators, unable to self-train quickly, reverted to simpler tools such as spreadsheets.

Optimizing data visualization for human attention, especially in the early phases of data exploration and data triage, has emerged as a clear goal. To this end, we developed VizDeck, a method and prototype for web-based visual analytics over relational data that uses a card game metaphor and automatic visualization techniques to assist users in creating interactive visual dashboards with no programming. Informed by the statistical properties of the underlying dataset, VizDeck pre-generates candidate visualizations (and UI widgets) called vizlets, displaying them to the user in a grid. These vizlets can then be browsed, discarded, or promoted for inclusion in an interactive dashboard. The user does not need to perform explicit configurations to construct a dashboard; vizlets automatically respond to brushing and linking events triggered by other vizlets.

Figure 1. VizDeck automatically generates a “hand” of thumbnail visualizations called vizlets.

Dashboards can then be saved and shared among collaborators by exchanging URLs. Besides affording an undo/redo capability, this technique allows a collaborator to review the steps taken by the original author to create the dashboard. This type of provenance has been shown to improve cross-training and communication between users (Bavoil et al., 2005).

To assess how the process of automating the visualization process affected data exploration, we conducted an initial user study of 32 subjects (with 315 trials run) testing their ability to complete data analysis tasks. We also conducted semi-structured interviews with each of the subjects to better understand their sensemaking process. Russell et al. (1993) describe sensemaking as “the process of searching for a representation and encoding data in that representation to answer task-specific questions”, noting that this process requires varying cognitive and external resources. In our design and evaluation of VizDeck as a tool, we were particularly interested in understanding how both efficiency and sensemaking were affected by the automation of the visualization process.

The contributions of this work are 1) a new interface design that accelerates visualization selection and dashboard assembly tasks for researchers when compared to more complex visualization tools, 2) a quantitative and qualitative analysis of users’ rapid exploration of scientific data, and 3) a prototype implementation that is available on the web (http://escience.washington.edu/vizdeck).
Related Work

There have been previous examples of systems that attempt to capture the data exploration process. Jankun-Kelly et al. (2002) proposed a model for recording multiple visualizations as part of a research process. Similarly, VisTrails, a system developed by Bavoil et al. (2005), offers a framework for showing multiple visualization pipelines within the same visualization “spreadsheet”. While these tools provide ways to view multiple visualizations, they are designed for experienced users who expect a significant degree of control over the visualization, and are not optimized for the early stage data triage task.

More recent work on intelligent user interfaces attempts to infer the user’s task from behavior, and use this information to recommend visualizations (Gotz & Wen, 2009). Dörk et al. (2008) derive coordinated visualizations from web-based data sources. Additional frameworks for creating mashups or simple interactive applications combining data from multiple sources have been proposed (Elmeleegy, Ivan, Akkiraju, & Goodwin, 2008; Ennals & Garofalakis, 2007), assuming a pre-existing repository of mashup components. Elias and Bezerianos (2011) focus on dashboard creation by both experts and novices and include chart recommendation features. These relatively sophisticated dashboard components must be constructed individually by the user through a multi-step process. Other visualization tools such as Many Eyes (Viegas, Wattenberg, van Ham, Kriss, & McKeon, 2007) and Tableau’s “Show Me” feature (Mackinlay, Hanrahan, & Stolte, 2007), also take an approach of providing an existing menu of predefined visualization types for users to explore. In contrast, VizDeck generates and presents a diverse set of fully realized visualizations for users to browse and choose from.

Previous research points to the importance of evaluating visual analytics tools within the context of the users’ sensemaking process. Scholtz (2006) recognized the importance of going beyond usability as a metric for evaluating visual analytic systems, noting that time and completion should be augmented with other metrics that account for factors including situational awareness, interactions, and creativity. Kang et al. (2009) evaluated Jigsaw, a multi-view system for investigative analysis, and noted the importance of understanding the sensemaking process, especially as it related to visual exploration of data in a study comparing the tool to lower-tech methods, including paper and pencil.

Design Considerations

Objective assessment of visualization quality is considered one of the grand challenges in the visualization community (Johnson, 2004), and is a prerequisite for automatic visualization of data. Seminal approaches to this problem involve heuristics based on human perception (Mackinlay, 1986). In VizDeck, we encode this style of heuristics (when they exist) to influence the recommendations, but augment them by training a model that learns the relationship between statistical features of the data and visualization properties. In this paper, we address the effect of the UI design features on usability and sensemaking rather than focusing on system design or the details of the statistical recommendations. In designing VizDeck to afford rapid exploration of scientific datasets, we took into account 1) Tufte’s (1990) principle of small multiples exploration, 2) automatic brushing and linking, and 3) the importance of human-in-the-loop interactions.

Small Multiples Exploration

We wanted to provide an efficient manner for users to pan the visual field of the dashboard display and detect trends or surprising results in the dataset. These thumbnails, which we refer to as vizlets (broadly defined as any UI element that either displays data or accepts user input), were inspired by small multiples (Tufte, 1990). Tufte argued that displaying several simpler visualizations can be more effective than displaying one dense, data-rich visualization. The VizDeck interface uses vizlets to compactly express the structure of many visualizations at once. The user rearranges these vizlets, discards those that are irrelevant, and promotes those that warrant closer inspection (Figure 1). We liken this process to a card player manipulating their hand in preparation for strategic play. To support efficient browsing of many screenfuls of vizlets, we also provide a search-box feature to allow users to inspect groups of related vizlets at once. Search terms may include references to vizlet type such as “scatter” or “line,” or references to attribute names, such as “oxygen” or “fluorescence” (Figure 3). We provide basic auto-completion for the search bar with disjunctive semantics (term1 OR term2 OR term3) to decrease the likelihood of empty result sets.
Automatic Brushing and Linking

Once promoted, vizlets typically expose additional interactivity features. For example, values can be selected in drop down boxes, and points on the scatter plots can be highlighted. These interactions automatically affect the display of all other vizlets, both those in the dashboard and those in the grid. Unlike similar features in other tools (e.g., Tableau), no explicit configuration is necessary to establish links between vizlets. As a result, users can assemble coordinated interactive dashboards in just a few seconds with no programming and no explicit configuration; Figure 2 illustrates an example of this feature. The underlying dataset arises from measurements from an environmental flow cytometer called SeaFlow (2011), an instrument deployed on oceangoing vessels that automatically measures population counts of microorganisms. The user has selected a drop down box displaying a list of microorganism types (upper left), a timeline of total particle counts by time (lower left), a histogram of population counts for each type by hour (upper right), and a scatter plot of two optical properties (lower right). Each vizlet must respond to this action, but the exact semantics of these events vary depending on the vizlet type: designers are free to define the specific behaviours of the vizlets as they see fit.

Figure 2. An interactive dashboard in VizDeck where the user has added four vizlets to the dashboard.

Figure 3. Keyword search equipped with auto-completion can be used to filter the grid of vizlets.
Human in the Loop Interactions

As we have described, the vizlet designers have some control over how their visualizations will be displayed and when they will be recommended. However, we emphasize human-in-the-loop interactions to improve the recommendations over time, independently of the static, design-time considerations. In particular, we log the interactions of the user and use the results to inform a decision tree-based ranking function such that the system learns over time which types of vizlets are more likely to be effective for a dataset with particular features.

Users either promote or discard vizlets based on the thumbnail’s perceived utility for their task. When a user promotes a thumbnail vizlet to a full-sized vizlet, we interpret that as a “like” vote from the user, and a discard action as a “dislike” vote from the user. We describe each vizlet with a feature vector and use it to train a linear model of historical voting data to assign a score to the candidate vizlets with respect to the underlying data. We are currently addressing appropriately weighting these human-in-the-loop interactions with more structured ranking heuristics based on display type. These details are largely independent of the UI methodology we evaluate in this paper, and are not discussed in detail.

Implementation

The VizDeck system is comprised of client and server subsystems. The server is written in Ruby and is responsible for connecting to data sources and retrieving results of queries presented by the user on the client side. The JavaScript client runs in the browser and is responsible for rendering the vizlets and controlling user interactions. VizDeck currently consumes data from SQLShare, a database-as-a-service application that favors ad hoc integration, query, sharing, and visualization over predefined schemas (Howe et al., 2011). Each dashboard is automatically assigned a unique, permanent URL when first created. This URL can be shared with other users at any point in the authoring process allowing researchers to share interesting findings.

When the server connects to a data source and retrieves results for a query, it creates compact definitions of vizlets called protovizlets. A protovizlet is a set of key/value pairs that collectively describe the visualization or UI widget so it can be rendered on the client side. At a minimum, the protovizlet must contain a “title” (a string displayed to the user) and a “type” (a string indicating the rendering method to use on the client; e.g., scatter, bar, etc). In the current system, we offer a variety of simple vizlet types: bar charts, histograms, line charts, timeline plots, pie charts, maps, and multi-select boxes (for selecting values to filter the data displayed in other vizlets). The client then converts these protovizlets into interactive vizlets that are displayed to the user. Before the protovizlets are sent to the client, the server examines the scores and sorts the protovizlets by their score. The score is computed by estimating the likelihood of selection of each vizlet based on a learned relationship between statistical features of the data used to generate the vizlet and the visualization type. To train the model, we use log data collected by the system as ground truth. Each time a user promotes a vizlet, we interpret the promotion as a vote that the vizlet was considered important enough to inspect more closely. Similarly, a discard action is a vote that the vizlet was not useful to the user. A positive net sum of these votes across users indicates the classification as “good” or “bad.” We then attempt to predict this classification from features of the data. For example, a line chart is less useful for sparse and irregular data, and a bar chart is not useful when there are too many or too few categories. The details of the scoring algorithm are outside the scope for this paper. The server also compares the protovizlets to each other with similarity scores to diversify the order in which the vizlets are to be presented to the user. Finally, the server selects the top k protovizlets to recommend to the user.

Evaluation Methodology

We conducted a quantitative and qualitative evaluation to assess VizDeck’s performance for finding relevant information within a scientific dataset when compared to three commonly-used data visualization tools. We hypothesized that subjects would be able to compare, analyze, and explore information within the dataset faster with VizDeck’s automated approach. The study included 32 participants (14 female), ranging in occupation and ages (19 to 60). Subjects were recruited through campus mailing lists. Sixty-six percent of the subjects were from science, engineering, or medical fields, while the others came from social science, business, education, and public policy. As data analytics
becomes an increasingly pervasive need among various research communities, we decided we did not want to limit participant involvement to a particular field. Almost seventy percent of the subjects identified their primary occupation as students. Thirteen of the participants reported having moderate to high levels of experience with data visualization tools, with the most commonly reported methods of visualization being Excel, R, and MATLAB.

We evaluated VizDeck along with three additional interactive visualization software tools: ManyEyes, Google Fusion Tables, and Tableau. We selected these particular tools on the basis of their functional similarity to VizDeck and ability to rapidly visualize data with no programming necessary. Many researchers employ different tools depending on their domain and require training or development skills; we distinguish our focus from that of high performance, 3D scientific visualization tools such as ParaView (Aherns, Geveci, & Law, 2005). We were primarily interested in exploratory data analysis and “data triage” tasks that emphasize “rapid prototyping” of visualizations rather than in-depth analysis. We therefore excluded tools such as IDL and MATLAB from our study because they are programming environments rather than end-user applications. We wanted to better understand how scientists and researchers who do not necessarily have a programming background can quickly and easily make discoveries in their data. We also did not test against Microsoft Excel since we were interested in tools that offered interactive visualization functionalities with varying degrees of automation.

The three tools evaluated within the user study were:

Tableau Desktop: Tableau is a highly polished, feature-rich commercial visualization tool that allows users to drag and drop field names to visualize data and create interactive dashboards. The desktop version of the software runs on Windows and can open data files from a variety of Windows compatible formats. It also provides a “Show Me” feature, a set of UI commands for automatic presentation (Mackinlay et al., 2007).

Google Fusion Tables: Fusion Tables is a Google web application that allows users to visualize and publish data as maps, timelines and charts. Data tables can be uploaded from spreadsheets, CSV, or KML format and the tables are hosted online (Gonzalez, 2010).

Many Eyes: Many Eyes is an online data visualization tool produced by IBM Research and the IBM Cognos Group. It allows users to cut and paste data into a field and then select visualization types and share these with others. The goal of the site is to promote a “social style of data analysis” in which visualizations promote discussions among users (Viegas et al., 2007).

Experimental Design

We conducted a user study with 32 participants each performing tasks on two of the four interfaces (a software error excluded one set of trials). The study included both between-subjects and within-subjects variables, since each subject evaluated two software interfaces. Subjects were asked to utilize one of the data visualization tools to complete five tasks regarding data about the periodic table of elements. The data set included information on element names, atomic number, boiling points, melting points, first-ionization energies, group number and period number. Properties included those of chemical elements from Hydrogen (H, atomic number 1) to Uranium (U, atomic number 92). For each element, available data included: atomic number, boiling and melting points in Kelvin (K), first-ionization energy in electronvolts (eV), period (1-7), and group (1-17). These are basic chemical properties that are taught in undergraduate chemistry coursework and considered important to understand relationships and chemical properties of the elements.

The performance of each subject was evaluated for the use of two of the software interfaces. In total 15-16 subjects performed the five questions on each interface. Each task was video recorded for researcher analysis, and VizDeck interactions were logged by the software system. Subjects were explicitly not given a training tutorial on any of the systems with the underlying assumption that training sessions would take valuable time away from researchers and scientists needing to perform science or analyze data. Subjects were informed that they could use the help documentation located within the software tool if necessary, although it was explained that time was a determining factor in evaluating how well they performed each task. The experimenter did not answer any direct questions regarding the capabilities of features of the interface, only noting that each question was in fact answerable using the given software interface. After subjects had completed the computer tasks, they were asked to describe the ease of use and their experience with each interface. This method was chosen over methods where
users might explain their thought process during the task itself in order to provide a more realistic estimate of time on task. This study was conducted in a university lab setting using desktop computers.

**Task Description**

The tasks were chosen to represent several types of visual data exploration tasks that could be answered with a brief inspection of a simple visualization: identifying a trend in a single variable, finding the mean, comparing two properties, finding the highest value, and recognizing a visual relationship between properties. These interactions were chosen as important qualifiers based on heuristics for presenting data for visual quantitative analysis (Few, 2009). The task order was randomized to account for learning effects, and two sets (Set A and Set B) of nearly identical tasks were alternated so that the answers were different between the interfaces.

The evaluation tasks for both Set A and Set B were as follows:

**Task 1. Identify a Trend within a Group.** Is there a noticeable trend for melting points (Set A) or boiling points (Set B) of the elements in Group 17 (Set A) or Group 12 (Set B)?

**Task 2. Find the Average for a Given Period:** Identify the correct average first-ionization energy for Period 3 (Set A), or correct average first-ionization energy for Period 6 (Set B).

**Task 3. Compare Two Properties.** Compare the melting points and first-ionization energies for Group 6 elements (Set A), or boiling points and first-ionization energies for Group 17 elements (Set B), and identify whether both properties were increasing, one was increasing while the other was decreasing, or both were decreasing.

**Task 4. Find the Highest Value.** Find the element with the highest boiling point (Set A) or the highest melting point (Set B) out of a collection of four elements. Finding the highest value in a list of numbers would not typically require visualization; we included this question primarily as a control.

**Task 5. Fit a Curve or Line.** Find a relationship between boiling point and some additional property that visually matched a given curve (Set A) or line (Set B). This question was unique in that it had users consider a visual representation of the data. We intended to capture situations where researchers might anticipate certain relationships in the data but are uncertain which variables will exhibit it. For example, clinical studies may generate hundreds of measurements about each patient and researchers may be looking for any measurements that visually follow the pattern of expression of a particular symptom.

**Evaluation and Results**

We found that the particular visualization tool used did have an impact on completion rate and time to completion for each task. The completion rate for the combined tasks shows that Fusion (97%) and VizDeck (94%) were comparable, with better performances than with Many Eyes (78%) and Tableau (68%), as shown in Figure 4 (left). Fusion (105 seconds) took on average the shortest amount of time to complete for all tasks combined, followed by VizDeck (130 seconds), Many Eyes (157 seconds) and Tableau (172 seconds) as shown in Figure 4 (right). We only included times on tasks that were successfully completed because we did not want to unduly weight the instances where subjects took a short amount of time but did not provide a correct answer. The withdrawal of these times made it challenging to establish statistical significance for time to complete. Therefore, we wanted a way to combine both the completion rate and the time on task to provide a more comprehensive and appropriate measure of efficiency.
Figure 4. Combined task percent completion with error bars calculated at a 95% Confidence Level using the Adjusted Wald Method for Confidence Intervals (left) and mean time to complete combined tasks (right) for each visualization tool.

The National Institute of Standards and Technology (2001) efficiency metric provides a combined score for completion success and time to complete. The metric is a ratio of the task completion rate and time to complete and the Common Industry Format (CIF) for Usability Test Reports describes this NIST measurement as the “core measure of efficiency” (NIST, 2001). The NIST ranked the software tools similarly to the completion percentage, as shown in Figure 4. One of the challenges of using the NIST efficiency metric is that it does not provide a way to reliably determine the statistical significance of the result since it uses the mean before combining the time and completion, and cannot be calculated for each trial run. Additionally, it requires a reasonable unit of scale for each task, making it harder to compare results across questions.

Figure 5. The NIST efficiency metric for combined tasks for each software tool.

Results: Time and Success Score

To determine the statistical significance of the results for each question we used the Time and Success Metric (TSM) as an efficiency metric that allows us to calculate a combined score for each individual trial (Perry, Aragon, Meier, Peffer, & Pritoni, 2001). The TSM takes time and the completion for each trial run as inputs, and is normalized between 0 and 1 by using a logistic transformation, offering a way to compare results across questions. The TSM score is given by the following equation:

\[ M_i = \frac{2s}{1 + e^{xi}} \]

where
The Time and Success Metric score was calculated for each question, showing variability in ranking score dependent on the question (Figure 6). The data are not normally distributed, so we used the non-parametric Kruskal-Wallis H test to determine statistical results. Table 1 shows the Chi-square value (H) and significance level for each task (Table 1). Pairwise comparisons for the Time and Success scores using Mann-Whitney U tests also indicated significant differences between the tools (Table 2).

Table 1

<table>
<thead>
<tr>
<th>Task</th>
<th>H</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1.</td>
<td>12.019</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>T2.</td>
<td>18.885</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>T3.</td>
<td>30.681</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>T4.</td>
<td>18.331</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>T5.</td>
<td>12.689</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Table 2

Mann-Whitney U Test comparing the Time and Success Scores for tools. The bold border indicates tests involving VizDeck. Shaded boxes indicate tests for which VizDeck’s performance was measurably better.

<table>
<thead>
<tr>
<th>Task</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion vs.</td>
<td>Fusion</td>
<td>no sig. diff.</td>
<td>Fusion</td>
<td>no sig. diff.</td>
<td>no sig. diff.</td>
</tr>
<tr>
<td>ManyEyes</td>
<td>U=59.5, P&lt;0.05</td>
<td></td>
<td>U=7, P&lt;0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fusion vs.</td>
<td>Fusion</td>
<td>no sig. diff.</td>
<td>Fusion</td>
<td>no sig. diff.</td>
<td>Fusion</td>
</tr>
<tr>
<td>Tableau</td>
<td>U=56, P&lt;0.05</td>
<td></td>
<td>U=16, P&lt;0.01</td>
<td>U=58.5, P&lt;0.05</td>
<td>U=39.5, P&lt;0.01</td>
</tr>
<tr>
<td>VizDeck vs.</td>
<td>Fusion</td>
<td>no sig. diff.</td>
<td>VizDeck</td>
<td>Fusion</td>
<td>Fusion</td>
</tr>
<tr>
<td>Fusion</td>
<td>U=51, P&lt;0.01</td>
<td></td>
<td>U=59.5, P&lt;0.05</td>
<td>U=16, P&lt;0.01</td>
<td>no sig. diff.</td>
</tr>
<tr>
<td>VizDeck vs.</td>
<td>VizDeck</td>
<td>VizDeck</td>
<td>VizDeck</td>
<td>ManyEyes</td>
<td>no sig. diff.</td>
</tr>
<tr>
<td>ManyEyes</td>
<td>U=69.5, P&lt;0.05</td>
<td>U=34, P&lt;0.01</td>
<td>U=40.5, P&lt;0.01</td>
<td>U=63.5, P&lt;0.01</td>
<td></td>
</tr>
<tr>
<td>VizDeck vs.</td>
<td>VizDeck</td>
<td>VizDeck</td>
<td>VizDeck</td>
<td>no sig. diff.</td>
<td>VizDeck</td>
</tr>
<tr>
<td>Tableau</td>
<td>U=60, P&lt;0.05</td>
<td>U=38, P&lt;0.01</td>
<td>U=48, P&lt;0.01</td>
<td></td>
<td>U=53, P&lt;0.01</td>
</tr>
<tr>
<td>Tableau vs.</td>
<td>VizDeck</td>
<td>no sig. diff.</td>
<td>no sig. diff.</td>
<td>no sig. diff.</td>
<td>VizDeck</td>
</tr>
<tr>
<td>ManyEyes</td>
<td>U=60, P&lt;0.05</td>
<td></td>
<td></td>
<td>ManyEyes</td>
<td>U=90.5, P&lt;0.05</td>
</tr>
</tbody>
</table>

The Time and Success scores (Figure 6) show Fusion and VizDeck achieved better scores across all tasks, with the exception of VizDeck’s performance on Task 4 ("Find the Highest Value"). This task was specifically designed knowing the limitations with VizDeck. Results showed Fusion performed better at identifying a trend, comparing two properties, and sorting data from high to low, while VizDeck outscored the three other tools at finding an average. Overall VizDeck’s method of automated visualization outperformed Tableau and Many Eyes on the majority of tasks and attained comparable performance to Fusion on Tasks 1-3 and Task 5. Our goal was not only to analyze time and accuracy on tasks but also to better understand challenges and opportunities in users’ sensemaking processes in using these systems to facilitate rapid exploratory analysis of data. After gathering the quantitative results, we then turned to qualitative data to extract user sensemaking strategies.
Figure 6. The mean Time and Success score with 95% Confidence Levels for each task where higher (the closer to 1) is better performance. Fusion Tables results are attributable in part to the subjects’ use of non-visualization features that are incomparable to the other systems.

Discussion

While the five tasks as quantitative measurements give indicators of software performance, users’ qualitative evaluations and visualization tactics provide a way to interpret the results within the context of the sensemaking process. Participant interview responses and comments, along with analysis of the video and log data, provide a further context for understanding users’ investigative and analytic strategies that are described below.

Catch and Release. One commonly used strategy for analyzing the visual data presented in VizDeck was what we have termed a “catch and release” tactic, in which users promote one or more vizlet thumbnails to the dashboard to inspect them, and then, if they did not contain relevant information, “release” them back into the display area below the dashboard by clicking the discard icon. While these select and discard interactions provided helpful data for the human-in-the-loop interactions, this strategy actually cost users valuable time in the search process.

One seeming reason for such a strategy was users’ reliance on the vizlet titles to provide a context for the thumbnail view. By first promoting the vizlet to the dashboard, users could gain a clearer view of the title, and decide whether the visualization contained relevant data. One participant identified the challenge of gaining cues from the titl es in the thumbnail view noting, “It was easy to use to drag and drop [vizlets to the dashboard], but it was harder to read the titles, and sometimes they weren’t visible.”

Research on information scent processes provides the context of information foraging (Pirolli, Card, & Van Der Wege, 2001) as a way to better understand this catch and release strategy. VizDeck automated much of the work for the user in creating the visualization; yet conversely, it increased the need to provide clearly visible and explicit information scents, or cues, for each visualization thumbnail. Further exploration is needed to determine information scent cues that could most effectively allow users to scan the thumbnails without the additional time taken evaluating the image in a dashboard view. One possible method of overcoming this might be to provide an enlargement of the vizlet upon mouseover. To further support a strategy like catch and release, it is critical that visual interfaces provide at a glance metadata about the visualization.

Thinking Ahead of the Visualization. Another common strategy observed in the tasks was users’ attempts to guess the visualization type that would be most relevant to the given task. This was most evident in interactions with Many Eyes, as users had to select different visualization types during their initial interaction with the interface. Users therefore spent large amounts of time switching between visualization types in their attempts to find the right one. One user of Many Eyes went back and forth between the different visualization types (histogram, bar, scatter plot, etc.) eleven times before settling on one. A subject who was a software trainer by occupation, noted the challenge of having to guess which visualization type would be most useful, saying “[With Many Eyes] I have to think ahead of the data about
what I would like." This challenge highlights one of the key benefits of VizDeck’s automated approach in pre-creating visualizations of a variety of different types based on the properties of the data. VizDeck notably reduced the “visualization choice” problem to one of a “search-and-browse” problem.

**Table Filtering.** A common strategy employed by subjects using Fusion Tables was the use of the table view to scan and order the data. It should be noted that while both Many Eyes and Tableau afforded a view of the raw data in table format, users that encountered this view rarely used it to complete tasks, instead going back to the visualizations to answer questions. Conversely, there were numerous Fusion Tables users who attempted to complete tasks using one of the visualization views, and then reverted back to the table view.

User comments provide insights into the sensemaking processes that led users to use the table view instead of visualizations. One user commented that it “felt more instinctual, primarily because I have experience with Excel and Access,” a sentiment that was echoed by several users. Another user stated, “it was easy to sort through things.” Two considerations arise from such comments: one is that the table view was indeed more familiar to many users than the visualizations, due to their experience with Excel, and second is that the ability to manipulate and “sort” through the data (which Fusion afforded and the other tools’ table views did not) was key to Fusion Tables’ successful use. The latter observation was corroborated by the several failed attempts of users who tried to use the raw data view in Many Eyes and Tableau. In these tools, because the raw data was static, and had no affordances for sorting or filtering the data, users found it challenging to find the information they needed. Raw data browsing of large datasets (GB and TB) are notably difficult to inspect, which is why we had focused our evaluation on tools that allowed for rapid visualization. However, users’ familiarity with filtering data within Microsoft Excel raises compelling questions as to how similar functionality could be supported within an interface that scales with larger datasets.

**Visual Data Insecurity.** Another interesting strategy that emerged from our observations was a sense of insecurity that many users displayed regarding their selection of the correct visualization. Subjects using Fusion Tables would often cycle through several iterations of visualizations (changing the properties of the x-axis and y-axis) before selecting a visualization to answer a question. This often meant that users scanned and passed over relevant data in their quest for a better visualization. This process was similarly observed in VizDeck users for the “Fitting the Curve” task. When looking at the outliers of the time-to-complete for VizDeck for this particular task, two (of the three outlier) subjects actually selected the correct visualization, but continued to promote several additional vizlets to the dashboard. These actions suggest that users were insecure about their first visual instinct, even when it had been accurate.

The majority of the Fusion users rapidly applied filters, and were able to cycle through all options until there were no longer any combinations of properties left to view. This strategy seemed to provide an “end of the line” assurance and additional security to the users, in which they would then go back to the most relevant visualization created. Such observations provide insights into a much deeper challenge: how can visual analytic interfaces be built to support a sense of security and assurance to users that they have indeed checked the most relevant visual options? We posit that tools should be designed to provide easy means of supporting this verification process and not just the discovery process.

One of the primary issues that many subjects identified was the trade-off between ease-of-use in automating the visualization creation process and the control in manipulating these visualizations. Many users expressed the opinions that VizDeck was easy to learn, and quick to view. However, some users commented on the lack of power in customizing what they were viewing. One user noted “the learning curve for [VizDeck] is not as steep” as that for Fusion Tables, but that with Fusion Tables “I feel like I have more control over it.” One participant, a researcher with extensive experience using PyLab and MATLAB, noted a preference for greater analytic control in his evaluation of VizDeck, commenting that “the graphical web based interface is convenient but fees shallow for bigger more complex problems.”

Conversely, there were numerous subjects who found the simplicity of the automated “instant-view” thumbnails of VizDeck familiar to other graphical interfaces they had used. A chemistry student noted that VizDeck “feels like it was made for my generation, quick to observe the trends, and you can search for whatever you want.” This trade-off over how much the system or the human does the work in creating the visualization was a key motivator in the design and development of VizDeck. We were keenly aware that this was not going to be a tool designed for fine-grained visualization or complex function control. The user task “Find the Average Period” provides the strongest example of VizDeck’s successful use of automation (with an average completion time that was 67% faster than that of any tool). In this
task, it was clear that the human had to do less work to view and compare data on averages than they did with the other systems.

A “sort” option has not been integrated into the VizDeck tool, resulting in users’ poor performance on the “Find the Highest Value” task. This limitation in rearranging the data meant that users had to scroll through a chart, and we anticipated that VizDeck was unlikely to support this task. One option is to provide improved sorting functionality once the user has selected one of the automated vizlets. This challenge once again highlights the tension between the work required in creating the visualizations and the users’ sense of control in manipulating the data.

Future Work

There are several limitations in the VizDeck tool we would like to address in future work. We intend to refine the ranking function, which will become stronger in classifying vizlets as we gather data from the human-in-the-loop processes. We would also like to add additional vizlet types to offer the most appropriate visual cues for users to explore the data. Additionally, we plan on modifying VizDeck to better understand how much control should be provided in functionality to sort and organize data along several different categories, such as by high/low. Further work is also needed to connect VizDeck to other data repositories such as Google Fusion Tables and Microsoft Data Market. As data is increasingly shared among researchers in varying and diverse domains, we decided not to limit the background or experience of the subjects participating in the study. However, we do see the benefit of conducting a longer-term evaluation of VizDeck within the context of a particular domain and a much larger dataset to gain greater insights into the process of sensemaking in an applied context.

Conclusion

We introduced VizDeck, a novel web-based visualization tool for relational data that uses a card game metaphor and automatic visualization techniques to aid researchers in exploring datasets. We described the system implementation that allows the visualization thumbnails (or “vizlets”) to be produced and ranked within the dashboard. We then discussed an initial evaluation of VizDeck and three additional data visualization tools to gain greater quantitative and qualitative insights into users’ ability to navigate, compare, and quickly make decisions about data trends. Our quantitative evaluation found an automated approach particularly effective for aggregating results, comparing trends, and visually exploring a scientific dataset. A qualitative analysis yielded evidence of four types of user sensemaking patterns in rapid analysis of scientific data and demonstrated the viability of increasing the level of automation in visual analytics systems.

References


Citizens’ Use of Twitter in Political Information Sharing in South Korea

Jisue Lee
Florida State University
jl10n@my.fsu.edu

Hohyon Ryu
University of Texas, Austin
hohyon@utexas.edu

Lorri Mon
Florida State University
Lorraine.Mon@cci.fsu.edu

Sung Jae Park
Hansung University
spark@hansung.ac.kr

Abstract

This study examined citizens’ use of social networking site Twitter in political information sharing in South Korea. Content analysis was used in classifying message types and sentiments from the most frequently re-tweeted (RT) messages including the names of three top political leaders running for general elections in 2012. Correlation analysis comparing citizens’ use of Twitter in political information sharing online with results of public opinion polls offline indicated: 1) the volume and magnitude of re-tweeted messages are significantly correlated with results of public opinion polls; 2) types of messages are not correlated with the public opinion polling results; 3) positive and negative sentiment revealed in Twitter messages are highly correlated with the results of public opinion polls. Findings from this case study provide insights into citizens’ use of Twitter in political communication.

Keywords: political communication, information sharing, re-tweeting (RT), message type classification, sentiment identification

Problem Statement

Social media tools such as Twitter, Facebook and YouTube are now considered as politically transformative communication technologies as radio and television. There are predictions that social networking sites (SNSs) such as Facebook and Twitter will transform democracy, allowing citizens and politicians to communicate, connect and interact in ways never before thought possible (Grant, Moon, & Busby Grant, 2010). In Barack Obama’s Presidential campaigns in 2008 and 2012, over 100 staff members worked on Twitter outreach alone (@barackobama) (Sweet to Tweet, 2010; Campaigns Use, 2012). Current studies also show that the number of Japanese politicians using Twitter grew from three to 485 in under a year while in Germany, 577 politicians opened Twitter accounts (Hong & Nadler, 2011). Increasingly, politicians and elected officials are realizing the power of social media for communicating political information and interacting with citizens.

In considering the impact of social media in the political sphere, many researchers have explored how using SNSs such as Facebook and Twitter influences elections and public opinion poll results (Robertson, Vatrapu, & Medina, 2009; Hong & Nadler, 2011; Tumasjan, Sprenger, Sandner, & Welpe, 2010; O’Conner, Balasubramanyan, Rouledge, & Smith, 2010). These studies examine cases from the United States and European countries including Germany and the Netherlands. However, few studies from Asian countries were conducted. A case study on South Korea reviewed the evolution of hyperlinked networks, but did not explore use of Twitter during elections (Hsu & Park, 2011). Findings and insights from empirical studies for different countries are needed to understand how citizens and politicians worldwide share political information and opinion via social media. This study investigates in particular political information sharing in South Korea using Twitter.


Copyright is held by the authors.
Research Purpose

The purposes of conducting this study are twofold: to explore re-tweeting (RT) information behavior for political messages on Twitter by citizens in South Korea, and to compare the number and types of re-tweeted messages, and the sentiments captured from these messages with the results of public opinion polling about leading political figures. This research allows us to better understand the role of Twitter within citizens’ political information sharing in South Korea, and offers insights into relationships between the message types and citizens’ sentiments as expressed on Twitter and in public opinion polls.

Significance of Study

2012 was a significant year in South Korea. With a general election in April 2012 and the presidential election in December 2012, change of regime issues engaged citizens in South Korea. The National Election Commission in South Korea lifted a strict ban on using social networking sites (SNSs) such as Facebook and Twitter for election campaigns on the basis that using social media could broaden citizens’ obtaining and sharing of information on candidates and elections at lower cost. With increased enthusiasm regarding this change in South Korea, the April 2012 election was considered as an important testbed for the role of social networking sites (SNSs) in political communication.

This paper examines how citizens in South Korea used Twitter in sharing political information and opinions on three candidates running in the April 2012 general election, and investigates relationships between citizens’ use of Twitter and the results of public opinion polls. This study investigates South Korean citizens’ re-tweeting of political messages about three candidates in the April 2012 election as community behavior indicating political sentiments, agreement, and consensus of political opinion. The study compares results of public opinion polling with citizens’ re-tweeting of political messages in Twitter to provide insights into political information sharing in South Korea, and to add empirical findings in the growing body of research on social media use in political communication.

Backgrounds of Major Concepts and Literatures

Themes and research questions in this study concern the major concepts of Twitter as a social media tool, Diffusion of Innovation (DOI) theory and Re-tweeting (RT) in Twitter, and Use of Twitter and impact of on elections. Each will be examined for the further discussion.

Twitter as a Social Media Tool

Twitter, created in March 2006 and officially launched in July 2006, is a fast growing real-time social media tool allowing people to find and share information on what is happening worldwide (Chang, 2010). Twitter defines its service as “a real-time information network that connects you to the latest stories, idea, opinions and news.” (Twitter, 2012). By January 2011, Twitter had over 200 million users, and by October 2011 was handling over 350 million tweets per day (Roosevelt, 2012; Twitter launched, 2011). Twitter’s micro-blogging and messaging functionality has become a powerful tool for interpersonal, professional and academic communication (Java et al., 2007; Thomas, 2010; Dann, 2010).

Twitter messages allow a maximum length of 140 characters, and average 11 words per message (O’Connor et al., 2010). Messages, known as “tweets,” can be made public or hidden, directed at another user by including the “@” symbol followed by another user’s account name, i.e. @Friend_Username. Users can also share others’ messages by “re-tweeting” (RT) them; which copies and disseminates the original message to the user’s followers (Zhao & Rosson, 2009). Any message can be annotated with a topic or subject using hashtags, i.e. #Topic; clicking on or searching on a hashtag displays a choice of top tweets or all current tweets on Twitter that share the same hashtag. However, Twitter hashtags still suffer from their fragmentary and redundant nature (Chang, 2010). Therefore, this study excluded hashtag keywords (#Keywords) in the data collection process.

Diffusion of Innovation (DOI) Theory and Re-Tweeting (RT) in Twitter

Rogers (1995) defines diffusion of innovation (DOI) as “the process by which an innovation is communicated through certain channels over time among the members of a social system” (p.5). For this
study, DOI provides a strong theoretical background to explain the phenomenon of adoption of innovation of political information seeking and sharing via re-tweeting (RT) in Twitter.

According to Rogers (1995), an innovation can be any "idea, practice, or object that is perceived as new by an individual or other unit of adoption" (p. 11). The diffusion process includes four key elements: innovation, the social system that the innovation affects, the communication channels of that social system, and time (Rogers 1962, 1995). The notion of innovation in DOI has also been expanded to include new products, ideas, services, methods, and inventions (Chang, 2010). In this study, the innovation diffused among users particularly refers to the idea of seeking and sharing political information and opinions in Twitter through re-tweeting (RT).

Communication is defined as "the process in which participants create and share information with one another in order to reach a mutual understanding" (Rogers, 1995, p. 17). Mass media and interpersonal communications are two major communication channels in the dissemination process of innovation (Rogers, 1962, 1995). Growing use of Twitter through creating and re-tweeting Twitter messages on computers and mobile devices can be seen as a major new pattern of mass-communication (Zhao & Rosson, 2009). Dann (2010) emphasizes that Twitter has rapidly "evolved through user innovations with the re-tweet (RT), reply (@) and hashtag (#) marks being introduced by consensus and community behavior" (p. 1). Roosevelt (2012) also mentions that re-tweeted messages should be analyzed with weights since re-tweeting a message can be interpreted as agreement with that particular message, and can also spread the message faster and influence more users. Therefore, it is important to investigate the users’ re-tweeting (RT) as community behavior of agreement and consensus in this study.

Rogers (1995) defines a social system as "a set of interrelated units engaged in joint problem solving to accomplish a common goal" and further denoted characteristics of social systems as: social norms, opinion leaders, change agents, and types of innovation decisions, which can promote or hinder the diffusion of innovations (p. 23). The time aspect is essential for explaining the innovation-decision process, the impact of innovators on adopters and the growth rates of adoptions (Chang, 2010; Rogers, 1962, 1995). In this study, the time aspect offers findings on offline public opinion polling results in comparison with online users’ concurrent re-tweeting of political messages for a comparative view of reflected political opinion over time.

Use of Twitter and Impact on Elections

Many researchers have studied how use of particular social networking sites (SNSs) by politicians and citizens relates to results of public opinion polls and elections (Hong & Nadler, 2011; Tumasjan et al., 2010; O’Conner et al., 2010; Vergeer et al., 2011).

Tumasjan et al. (2010) argue that Twitter message content reflects the offline political landscape, thus potentially predicting actual election results. In a German case study, numbers of tweeted messages were observed to closely match ranking by share of the vote in election results, and nearly approximated results of traditional election polling. O’Conner et al. (2010) observed sentiments in Twitter messages replicated 2008-2009 U.S. consumer confidence and presidential job approval polls. Hong and Nadler (2011) studied U.S. politicians’ use of Twitter and its impact on public opinion, finding that the impact of the number of tweets was not significant for any of the tested opinion polls. Vergeer et al. (2011) explored the relationship between using Twitter and gaining votes in the Netherlands. Although the study showed a positive relationship between the number of Twitter messages and the number of votes, the size of the Twitter network was noted to be a limited indicator for voting outcomes.

Although there is some research on Twitter in political communication and elections in the U.S., Germany, and the Netherlands, as yet little research has explored political use of Twitter within Asian countries. While studies have examined Twitter messages created by politicians and citizens which include particular keywords, there has not been research particularly examining re-tweeted messages.

Therefore, this research studying re-tweeted messages in political communication in South Korea provides new insights into citizens’ use of Twitter in the context of the national elections in South Korea.

Research Questions

The goal of this study is to investigate relationships between citizens’ use of Twitter and public opinion polls, and to answer questions on how re-tweeting (RT) behavior in Twitter political communication may relate to public opinion polls.
RQ 1. Does the number of citizens’ re-tweeted messages correlate with results of concurrent public polls?
RQ 2. Do the types of citizens’ re-tweeted messages correlate with the results of public opinion polls?
RQ 3. Do the sentiments of citizens’ re-tweeted messages correlate with the results of public opinion polls?

Methodology

The procedure for collecting re-tweeted messages used the Python Twitter API to collect and store re-tweeted messages in a designated database. Collected Twitter messages were limited to those including three selected keywords for a specific time frame: the names of three leading political figures in South Korea from April 2 to May 11, 2012. Among those, the most frequently re-tweeted 200 messages for each class of keywords were selected as data sets of 600 messages. For this study, a set of 120 messages (20%) including the top 40 messages for each keyword class per one week were created, and total of 720 messages for three keyword classes for six weeks were used for the analysis.

Data Capture and Storage

Python Twitter API was used to collect re-tweeted messages including three specific keywords related to the general election of April 11, 2012 in South Korea - the names of three leading political figures: Bak GeunHye, Moon Jaeln, and Ahn ChulSoo (hereafter, Bak, Moon, and Ahn). Initially, re-tweeted messages including the hashtags #BakGeunHye, #MoonJaeIn, and #AhnChulSoo were collected. However, hashtags with the keywords showed very low usage among Korean citizens, and the fragmentary and redundant nature of hashtags as mentioned in Chang (2012) was also observed. Therefore, this study excluded hashtag keywords in data collection. The chosen keywords were appropriate in that these three leading political figures were expected to be the major candidates for the presidential election in December 2012.

The Python Twitter API collected tweets which included any one of the three presidential candidate names, automatically calculated frequency of messages being re-tweeted every 100 seconds, sorted the top 200 messages based on the re-tweeting frequency, and stored them in the designated database. As the Python Twitter API only included Twitter messages for the most recent six days, collected data needed to be stored in a different database.

Twitter messages were collected every 5 weekdays for 6 consecutive weeks, the same period as the collection of public opinion poll results. To study the relationships between Twitter messages and public opinion polling, correlation analysis was conducted. An example of Twitter data collected from April 2 to May 11, 2012 is shown below in Figure 1.

![Figure 1. Screen Capture of Collected Data](image-url)
Analysis

Re-tweeted messages including the three keywords (political leader names) were harvested from April 2 to May 11, 2012. A total of 556,675 messages including the keywords were re-tweeted. The most frequently re-tweeted 200 messages from each keyword class for each week were listed as a data set of 600 messages, totaling 3,600 messages for the three classes for 6 weeks. From among these, sets of 120 messages including the top 40 re-tweeted messages for the three classes weekly were accumulated for 6 weeks. A total of 720 re-tweeted messages were analyzed for this study.

Messages were analyzed using two types of content analysis: 1) classifying message types and 2) identifying the general sentiment expressed in the messages. The results of content analysis for re-tweeted messages were compared with the results of public opinion polls through correlation analysis using SPSS 18.

Classification of Types of Message

Twitter messages are created and re-tweeted with multiple purposes and motivations. The objective in classifying types of messages was to understand various purposes of citizens in their message-creating and re-tweeting behavior. Therefore, coding was conducted with a focus of studying the purposes as to why citizens created and re-tweeted particular messages.

Coding Scheme. The coding scheme was based on criteria from the literature. Pear Analytics (2009) categorized tweets as primarily related to pointless babble (40%), conversational (38%), pass along (9%), self-promotion (6%), spam (4) and news (4%). Content categories from Java et al.’s study (2007) included daily chatter on the daily routine of individual users, conversations which included replies to other users, information or URL sharing which were classified according to the presence of full length or shortened URLs, and news sharing which include sports, weather and commentary on current affairs. However, these studies classified types of Twitter messages, not types of purposes in creating messages. Robertson et al. (2009) examined linkage patterns of politically-oriented community networking on Facebook, classifying five types of linkage motivation patterns shown on three presidential candidates’ Facebook walls: evidence, rebuttal, action, joking and ridicule, and direct address. This study showed purposes for which posters created postings to share political information and opinions through social networking sites of Twitter.

In addition to adapting these five types of messaging from Robertson et al. (2009), three more types were created reflecting characteristics of the collected data set: Media Report, Human Report and Event Report. Detailed explanations on codes and definitions are shown in Table 1 (hereafter, E, R, A, JR, DA, MR, HR, ER).

Reliability Testing. Content analysis requires that researchers make “replicable and valid inferences from texts to the contexts of their use” (Krippendorff, 2004, p. 18). Methodological requirements of reliability and validity are critical demands in content analysis. This research undertook codebook development for content analysis through multiple intra-coder reliability tests and an interactive sequence of codebook revisions.

Intra-coder reliability. The initial coding scheme was based on Robertson et al.’s study (2009), adding additional codes emerging from the data to create a preliminary codebook. With the initial codebook, 3 rounds (6 iterations) of intra-coder reliability testing were conducted for three weeks. A sample set of 60 messages was coded to classify message types with a second coding was carried out of 60 messages one week later. The differences between the two trials of coding came from 19 out of 60 messages (31.6%). After revising the definitions of codes based on the analysis, a third coding with a different set of 60 messages was conducted. A week later, the fourth coding was conducted and differences between the third and fourth trial were discovered in 8 out of 60 messages (13.3%). With a re-revised codebook, a fifth coding with a different set of 60 messages was accomplished. In the following week, a sixth coding was completed with 100% agreement and no differences found between the coding results.
Table 1
Codebook of 8 Message Types

<table>
<thead>
<tr>
<th>Code</th>
<th>Sign</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence</td>
<td>E</td>
<td>Tweets written in order to <em>provide evidence for a particular opinion or simple fact</em>. It may include actual links to evidence such as newspaper, blog post, images, video and etc. along with the texts.</td>
</tr>
<tr>
<td>Rebuttal</td>
<td>R</td>
<td>Tweets written in order to provide negative responses or comments to rebut others’ ideas. It may sound like evidence, but it has additional components of <em>reaction or providing negative comments to specific persons or organizations</em>.</td>
</tr>
<tr>
<td>Action</td>
<td>A</td>
<td>Tweets written in order to <em>encourage others to take actions either on Internet such as participating in a poll, joining a group, or in the real world, such as donating money or attending a rally</em>.</td>
</tr>
<tr>
<td>Joking and Ridicule</td>
<td>JR</td>
<td>Tweets written in order to <em>ridicule something or someone, reveal something fully about them or their behavior, or simply point people to satirical content</em>.</td>
</tr>
<tr>
<td>Direct Address</td>
<td>DA</td>
<td>Tweets written in order to <em>directly address his/her own opinion or simply provide fact without any evidence</em>. It may include simple statements of his/her idea, fact, and opinion such as simple messages of support or lack of support as an expression.</td>
</tr>
<tr>
<td>Media Report</td>
<td>MR</td>
<td>Tweets written in order to <em>cite and report the contents from a Media Report</em>. It is usually followed by the actual links of the materials it refers, or can be written in the direct quotation form (&quot; &quot;). Even though the message itself includes neither links to the material it refers nor texts in direct quotation form, an indirect way of citation and reports of the contents from Media can be coded as MR. The main purpose of MR is to distribute and circulate the content of media.</td>
</tr>
<tr>
<td>Human Report</td>
<td>HR</td>
<td>Tweets written in order to <em>cite and report the contents from a Human’s remarks</em>. It is usually followed by the actual link of the material it refers, or can be written in the direct quotation mark (&quot; &quot;). Even though the message itself includes neither links to the material it refers nor texts in direct quotation form, an indirect way of citation and reports of the contents from Human remarks can be coded as HR. The main purpose of HR is to distribute and circulate the content of Human remarks.</td>
</tr>
<tr>
<td>Event Report</td>
<td>ER</td>
<td>Tweets written in order to <em>report and distribute the real-time event news, or his/her own experience of that event</em>. The main purpose of this ER is to distribute the facts or news, which happen as real time events but are not likely to be broadcasted by mainstream media.</td>
</tr>
</tbody>
</table>

The main reason noted to cause inconsistencies was that multiple purposes could be employed in creating a message. For example, a tweet could encourage people to vote by citing a politically important figure’s remarks. In this case, this message can be coded as Action (A) or Human Report (HR). Considering that ideas or opinions can be expressed using multiple methods including examples, rebutting, or making a joke, this ambiguity is intrinsic in interpreting human language itself. If one message includes characteristics of two types, it is possible that one main purpose could be understood better and stressed more over the other within the specific context so that each message can be coded as one specific type.

General Sentiment Identification

Tweeted messages include special features and conventions such as words, emoticons, and hashtags representing the author’s emotions and feelings within 140 characters (Agarwal et al., 2011; Tumasjan et al., 2010). These Twitter-specific features allow researchers to identify sentiments and opinions in tweeted messages. Sentiment analysis studies in Twitter are mainly conducted at the word level (Agawal et al., 2011; Tumasjan et al., 2010; O’Connor et al., 2010). To measure the emotional
meaning or degree in words and texts, researchers used pre-defined dictionaries such as the Dictionary of Affect in Language (DAL) or other systems with embedded dictionaries such as Linguistic Inquiry and Word Count (LIWC), or OpinionFinder, classifying words according to three categories: positive, negative and neutral. This approach enables researchers to measure sentiments at the word level, and research has indicated that the sentiment word frequency in tweeted messages correlates with the public opinion polling results (O’Connor et al., 2010).

However, it is still difficult to analyze sentiments or opinions from tweeted messages at the message level. Branthwaite and Patterson (2011) argue that “opinions are divided on the ability to code the underlying sentiments as positive or negative” and it is certainly difficult for machines to understand irony or sarcasm (p. 432). Twitter features such as emoticons and hashtags add variations, and thus sentiment analysis requires human intervention and qualitative analysis to increase validity and accuracy at the message level (Branthwaite & Patterson, 2011).

This study attempts to identify general sentiments understood and interpreted within political contexts. General sentiments within re-tweeted messages were sorted into three classes: positive, negative and neutral. Sentiments expressed toward the three leading political figures can be interpreted as an indicator of how Twitter users perceived those leaders. Overall sentiments were identified considering the context in the messages, not by particular words or terms. Intra-coder reliability was tested through the two trials of coding and agreement without error was relatively easily reached.

**Correlation with Public Opinion Polls**

One of the most frequently referenced public opinion polls in South Korea is Korean Gallup (http://www.gallup.co.kr). Korean Gallup polls are administered through telephone and mobile phone interviews to 1550 citizens every five weekdays. Favorable impressions of three leading political figures were collected from Korean Gallup polls for six weeks from April 2 to May 11, 2012. Correlation analysis was conducted to study the relationship between citizens’ expressed opinions online in Twitter and public opinion polling offline.

**Findings**

In answering the three research questions, general descriptive analyses including the magnitude of re-tweeted messages, classification of types of messages, and identification of sentiments were provided, with correlation analysis between results of content analyses of Twitter messages and the results of public opinion polls.

**RQ 1. Magnitude of Re-tweeted Messages and Public Opinion Polls**

A total of 556,675 re-tweeted messages including the three keywords were collected for six weeks - April 2 to May 11, 2012. The total number of re-tweeted messages including Bak for six weeks was 354,284 (63.64%), with 102,000 (18.32%) for Moon and 100,391 (18.03%) for Ahn (Figure 2). The number of top 40 most frequently re-tweeted messages including Bak for six weeks was 56,281 (53.69%), 24,880 (23.73%) for Moon and 23,674 (22.58%) for Ahn (Figure 3).

The magnitude of the re-tweeted messages including the three keywords varied, and might depend on public awareness of the three leaders. Bak is a leader of the ruling party of Saenuri, and a daughter of the former president during the military dictatorship from 1960s through 1970s, JungHee Bak. Moon is the leader of the biggest opposition party of United MinJoo, and is well known as the successor of the former president, MooHyun Roh. Ahn is a non-partisan university professor without any official experience as a politician. The magnitudes of both total re-tweeted messages and top 40 re-tweeted messages showed one pattern: Bak was most frequently talked about by Twitter users, followed by Moon and Ahn.
The extent to which the three leaders appeared in the Twitter timeline over the six weeks was not consistent with the percentages shown in public opinion polling results over the same period. Results from public opinion polls overall, as seen in Table 2, indicated that Bak obtained the most favorable impressions (averaging 36.83%) followed by Ahn (averaging 23%), and Moon (averaging 12.33%).

Table 2

<table>
<thead>
<tr>
<th></th>
<th>Week1</th>
<th>Week2</th>
<th>Week3</th>
<th>Week4</th>
<th>Week5</th>
<th>Week6</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAK</td>
<td>34</td>
<td>36</td>
<td>39</td>
<td>36</td>
<td>38</td>
<td>38</td>
<td>36.83 (%)</td>
</tr>
<tr>
<td>MOON</td>
<td>15</td>
<td>13</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>11</td>
<td>12.33 (%)</td>
</tr>
<tr>
<td>AHN</td>
<td>23</td>
<td>23</td>
<td>25</td>
<td>23</td>
<td>23</td>
<td>21</td>
<td>23 (%)</td>
</tr>
</tbody>
</table>

Notes. Public opinion polls results from nationwide data collection.

For correlation analysis, SPSS 18 was used. Total numbers of re-tweeted messages and of the top 40 most re-tweeted messages showed significant correlations with the percentages of public opinion polls (Table 3).
Table 3
Pearson Correlation between Re-tweeted Message and Public Opinion Poll

<table>
<thead>
<tr>
<th></th>
<th>RT_all</th>
<th>RT_40</th>
<th>Poll</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT_all</td>
<td>.985**</td>
<td>.676**</td>
<td></td>
</tr>
<tr>
<td>RT_40</td>
<td></td>
<td>.616**</td>
<td></td>
</tr>
</tbody>
</table>

Notes. RT_all = number of all Re-Tweeted messages; RT_40 = number of top 40 Re-Tweeted messages; Poll = Public Opinion Polls percentage.

** p < .01

RQ 2. Classification of Messages Types and Public Opinion Polls

A total of 720 messages were classified into eight types according to the codebook. Direct Address (326, 45.3%), Human Report (142, 19.7%), Joking/Ridicule (113, 15.7%), and Evidence (54, 7.5%) were the most frequently assigned types (635/720, 88.2%) across all the re-tweeted messages. The results from classification of each type were shown in Figure 4 and 5.

- The main purposes of creating and re-tweeting were to directly address opinions and information without evidence (DA), distribute the cited contents from human remarks (HR), make jokes or ridicule others’ ideas (JR) and provide evidence (E).
- Three categories, Media Report, Human Report, and Event Report (198/720, 27.5%), approximately one-fourth of messages, were created and re-tweeted to distribute and report cited contents from media, humans, and events. Media Report, Human Report, and Event Report represented the tendency of citizens to rely on other sources such as trustworthy media or leading figures’ remarks. Evidence (54/720, 7.5%) indicated messages created based on evidence. In sum, political messages identified with types of Evidence, Media Report, Human Report, and Event Report (252/720, 35%) taken together demonstrated the tendency of citizens to circulate Twitter messages including supporting evidence.
- Rebuttal and Joking/Ridicule messages provided negative responses, comments, or jokes about others’ ideas. Action messages sought to encourage people to take actions, and were usually created without including references or evidence. Direct Address messages directly address opinions and information without evidence. These four types of messages, Rebuttal, Action, Joking/Ridicule, and Direct Address (468/720, 65%), provided ideas or opinion without supporting evidence or references.
- Overall classification of message types found two major clusters with four types each: Supported (E, MR, HR, and ER; 35%) and Unsupported (R, A, JR, and DA; 65%). Supported message types included references to an information source such as a link, image, or remarks from particular media or figures, while unsupported message types did not.
Results from message type classification were shown in Table 4 and Figure 6. 240 tweeted messages for each political figure were analyzed.

- For Bak, Direct Address (114/240, 47.5%), Joking/Ridicule (50/240, 20.8%) and Evidence (27/240, 11.3%) were the most frequently classified message types. Supported message types (E, MR, HR, ER) were 65 (65/240, 27.1%) cases, while unsupported message types (DA, R, A, JR) were 175 (175/240, 72.9%) cases.

- For Moon, Direct Address (118/240, 49.2%), Joking/Ridicule (36/240, 15%) and Human Report (36/240, 15%) were the most frequently assigned types. As the senior advisor of the biggest opposition party against Saenuri Party, he is regarded as a progressive politician running against Bak. Supported message types (E, MR, HR, ER) were 81 (81/240, 33.7%) cases, while unsupported message types (DA, R, A, JR) were 159 (159/240, 66.3%) cases.

- For Ahn, Direct Address (94/240, 39.2%), Joking/Ridicule (27/240, 11.3%) and Human Report (86/240, 35.8%) were the most frequently assigned types. Unlike Bak and Moon, HR (86/240, 35.8%) was a dominant type for Ahn, and all of the HRs cite Ahn’s own remarks. This suggests that Ahn’s remarks were regarded as meaningful and influential among citizens, deserving repeated re-tweeting (RT). Even though unsupported types of DA, R, A, And JR (134/240, 55.8%) were observed more than supported types of E, MR, HR, and ER (106/240, 44.2%) cases, the contrast between two categories was more balanced than observed in the tweeted messages regarding Bak and Moon.
Table 4

Results from Classification of Message Types for Three Keywords

<table>
<thead>
<tr>
<th></th>
<th>DA</th>
<th>R</th>
<th>A</th>
<th>JR</th>
<th>E</th>
<th>MR</th>
<th>HR</th>
<th>ER</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bak W1</td>
<td>20</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W2</td>
<td>21</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W3</td>
<td>20</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>W4</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>W5</td>
<td>15</td>
<td>0</td>
<td>11</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>W6</td>
<td>20</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>Moon W1</td>
<td>18</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>40</td>
</tr>
<tr>
<td>W2</td>
<td>27</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td>W3</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>W4</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W5</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>14</td>
<td>7</td>
<td>40</td>
</tr>
<tr>
<td>W6</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>Ahn W1</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>25</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W2</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>21</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W3</td>
<td>16</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W4</td>
<td>15</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W5</td>
<td>15</td>
<td>5</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W6</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>Sum</td>
<td>326</td>
<td>16</td>
<td>13</td>
<td>113</td>
<td>54</td>
<td>35</td>
<td>142</td>
<td>21</td>
<td>720</td>
</tr>
<tr>
<td>%</td>
<td>45.3</td>
<td>2.2</td>
<td>1.8</td>
<td>15.7</td>
<td>7.5</td>
<td>4.9</td>
<td>19.7</td>
<td>2.9</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Notes. D = Direct Address; R = Rebuttal; A = Action; JR = Joking and Ridicule; E = Evidence; MR = Media Report; HR = Human Report; ER = Event Report. W1 means week 1 of 5 weekdays for which data collection is conducted.

Figure 6. Classification of Messages for Three Classes

The eight types of Twitter messages did not show any correlation with results from public opinion polls (Table 5). Thus, re-tweeted (RT) messages as classified by the purpose of the message did not relate to the public opinion poll results.
Table 5
Pearson Correlations Among Types of Message and Public Opinion

<table>
<thead>
<tr>
<th></th>
<th>DA</th>
<th>R</th>
<th>A</th>
<th>JR</th>
<th>E</th>
<th>MR</th>
<th>HR</th>
<th>ER</th>
<th>Poll</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>-1.65</td>
<td>.230</td>
<td>-274</td>
<td>.050</td>
<td>-1.87</td>
<td>-569*</td>
<td>-1.22</td>
<td>-0.23</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>-.104</td>
<td>.090</td>
<td>-1.81</td>
<td>.018</td>
<td>-.041</td>
<td>-.085</td>
<td>.074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.193</td>
<td>.666**</td>
<td>-.331</td>
<td>-.365</td>
<td>-.067</td>
<td>.232</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JR</td>
<td>-.015</td>
<td>.029</td>
<td>-.373</td>
<td>-.177</td>
<td>.237</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-.168</td>
<td>-.487*</td>
<td>.131</td>
<td>.233</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MR</td>
<td>.004</td>
<td>-.098</td>
<td>.214</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HR</td>
<td>-.084</td>
<td>-.233</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER</td>
<td>-.370</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poll</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes. DA = Direct Address; R = Rebuttal; A = Action; JR = Joking and Ridicule; E = Evidence; MR = Media Report; HR = Human Report; ER = Event Report.
* p < .05; ** p < .01.

RQ 3. General Sentiment Identification and Public Opinion Polls

Sentiments in tweeted messages were sorted into three classes: positive, negative and neutral. Sentiments were identified from the overall sentiment observed in messages, not by coding for particular terms, and are shown in Table 6 and Figure 7.
- For a total of 720 messages, 44.9% (323/720) negative sentiment, 46.1% (332/720) positive sentiment, and 9.0% (65/720) neutral sentiment were observed.
- For Bak, 98.3% (236/240) negative sentiment, and 1.67% (4/240) positive sentiment were identified. Tweets including Bak, an icon of conservatism, primarily showed negative sentiments.
- For Moon, 18.8% (45/240) negative sentiment, 75% (180/240) positive sentiment and 6.5% (15/240) neutral sentiment were observed. As a counterpart figure against Bak, tweeted messages showed mostly positive sentiments.
- For Ahn, 17.5% (42/240) negative sentiment, 61.7% (148/240) positive sentiment and 20.8% (50/240) neutral sentiment were found. Although overall sentiment toward Ahn was positive, a high portion of neutral sentiment was captured in tweeted messages.

Table 6
Results from Sentiments Identification

<table>
<thead>
<tr>
<th>Bak</th>
<th>NG</th>
<th>PS</th>
<th>NT</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>39</td>
<td>1</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W2</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W3</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W4</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W5</td>
<td>38</td>
<td>2</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W6</td>
<td>39</td>
<td>1</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>Moon</td>
<td>W1</td>
<td>8</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>W2</td>
<td>2</td>
<td>38</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>W3</td>
<td>8</td>
<td>30</td>
<td>2</td>
<td>40</td>
</tr>
</tbody>
</table>
Positive and negative sentiment observed in re-tweeted messages showed significant correlation with percentages in public opinion polling; however, neutral sentiment did not show any correlation (Table 7).

Table 7
Pearson Correlations Among Sentiments and Public Opinion

<table>
<thead>
<tr>
<th></th>
<th>NG</th>
<th>PS</th>
<th>NT</th>
<th>Poll</th>
</tr>
</thead>
<tbody>
<tr>
<td>NG</td>
<td>-.962**</td>
<td>-.555*</td>
<td>.842**</td>
<td></td>
</tr>
<tr>
<td>PS</td>
<td>.306</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td></td>
<td></td>
<td>-.869**</td>
<td></td>
</tr>
<tr>
<td>Poll</td>
<td></td>
<td></td>
<td></td>
<td>-.288</td>
</tr>
</tbody>
</table>

Notes.  G = Negative Sentiment; PS = Positive Sentiment; NT = Neutral Sentiment  
* p < .05; ** p < .01.
Conclusions

This study investigated political information sharing in social networking site of Twitter in South Korea. Re-tweeting as community behavior of agreement and consensus in Twitter was examined by both quantitative and qualitative approaches: counting the number (magnitude) of re-tweeted messages, using content analyses to classify types of messages and sentiments, and conducting correlation analysis between expressed opinions on Twitter and results of public opinion polls.

Although the magnitude of re-tweeted messages showed significant correlations with public opinion polling results over time, rankings of message magnitude differed from the rankings of public opinion polls. The magnitudes also appeared vulnerable to changes in political issues and events in real life. The 8 most frequently re-tweeted message types did not correlate with public opinion polling results. Contents in tweeted messages were highly subjective, complicated and contextual as a representation of political communication. However, sentiments in tweeted messages, while subjective and contextual as well, did show a correlation with public opinion polling results. This implies that capturing sentiments from tweeted messages dealing with broader political issues can be useful in gauging public opinion.

These findings require researchers to closely look at Twitter message content to understand purposes and underlying sentiments in context. In this perspective, the qualitative approach to classifying the purposes and sentiments of re-tweeted messages was appropriate in attempting to study this particular political communication phenomenon via social networking sites (SNSs). The significant correlations between the use of Twitter and public opinion polls indicate possible potential for utilizing current social networking sites (SNSs) to explore public opinion in political information sharing.

This study had several limitations. The sample used for this study may not be representative of South Korean electorate, as not everyone uses Twitter. Tweets collected for this study were also limited only to those including names of three top political leaders. Other specifically topic-based keywords such as titles of specific policy or political issues could be employed to collect targeted political data for future studies. This study also examined data for an abbreviated timeframe of six weeks. Therefore, further studies need to be conducted for a longer period of time in order to add more and stronger empirical findings.

References


When “Organizing Becomes an Extension of Your Brain”: DJs, Music Libraries and Information Practices

Jessica Lingel
Rutgers University
jlingel@eden.rutgers.edu

Abstract

As contemporary discourses of technology move from identifying information overload as a general concern to designing applications and conducting investigations that understand information overload on a more granular level, it becomes increasingly important to conceptualize what information overload means in everyday life. This paper addresses issues of being overwhelmed by media, looking specifically at a group of people for whom the stakes of media organization are quite high, namely DJs. With the emergence of digitally-stored music, music enthusiasts have been able to build increasingly large collections, to the extent that users frequently become overwhelmed by tasks of organizing their collections and creating usable playlists. Drawing on semi-structured interviews with 13 DJs, I examine obstacles in organizing digital music. With thick descriptions of how music technologies are and are not meeting the needs of a high-use group, this paper contributes to human information behavior research on organizational practices in everyday life within a particular community of practice.

Keywords: DJs, music, digital media, information practices, community of practice

Introduction

“There was like my parents’ definition of DJ, which was like Wolfman Jack, or whatever, and my definition of real DJ, which was Invisible Scratch Pickles and what not. And that has broadened both with technology, you know, now that there’s so many different ways to do it.” This description comes from Fin, a New York City based disc jockey (DJ), who is constructing a definition of DJ work, comparing his perception of what it means to entertain people with remixed and altered music with his parents’ definition of hosting a radio show. Interestingly, and central to the arguments I make in this paper, Fin argues that changes in technology have lead to changing definitions of what it means to be a DJ. But what, exactly, does Fin mean by this? What kinds of technology matter in defining DJ work? And what happens when those technologies change? This research project investigates the extent to which the definition of DJ work has (or has not) broadened as a result of technological change, focusing on experiences of information and media overload on the one hand and DIY (do it yourself) approaches to organizing and cataloging music on the other.

In order to expand understanding of existent practices of technology in everyday capacities, and to improve the design of systems related to music collection and production, I address a gap in research on actual information practices used to organize music in everyday life (Cunningham, Jones & Jones, 2004; Voida, Grinter, Ducheneaut, Edwards & Newman, 2005). In this study, I use ethnomethodological interviews with DJs to document organizational practices of media and analyze the impacts of technological change on a particular community of practice. Interviews with DJs enable inquiry into

Acknowledgements: Heartfelt thanks to participants for their time and feedback on this project. Thanks also to Danah Boyd, Nathan Graham, Ben Merriman and anonymous reviewers for their useful comments on this paper. Finally, many, many thanks to Stephen Bolles (aka DJ Still Life) for asking me for help organizing his music library in the fall of 2010, setting off my initial interest in this project, and being a vital resource of information and encouragement throughout the research process.


Copyright is held by the author.
practices of music organization among a high-use group. This in turn allows for critical analysis of other kinds of information management related to heterogeneous collections of digital media that are increasingly becoming integrated into the daily lives of groups as diverse as music lovers, academics (e.g. Brown, 1999; Palmer & Cragin, 2008) and amateur cooks (e.g. Hartel, 2006).

Prior Work

Within human computing interaction (HCI) literature, personal music collections have been studied both to improve design and to understand technology in everyday life. Regarding the former category, research on music and technology have been used to improve search functionality, such as visualization tools to assist users in managing their collections (Torrens, Hertzog, & Arcos, 2004). Among user studies in this area, Cunningham et al. (2004) analyzed a rich set of ethnographic accounts of personal music collections, focusing largely on organization of CDs and MP3s. Cunningham and Nichols (2009) explored music listening practices in small groups using participant observation and interviews. Voida et al. (2005) used qualitative interviews with a group of office employees to analyze practices of iTunes music sharing, and explored connections between reputation, identity and shared music collections. Benford, Tolmbie, Crabtree, Ahmed and Rodden (2012) studied collaboration between Irish folk musicians, revealing the extent to which practices of preparation were used even for work that was performed as impromptu. These studies contribute to a greater understanding of how technology and information are embedded in everyday music-related activities, such as parties (Cunningham & Nichols, 2009), live performances (Benford et al., 2012) and the office (Voida et al., 2005); In this project the emphasis is less on behaviors and norms within a given setting, and more on a specific kind of music collector – DJs. Because DJs rely on being able to tap their music collections in order to perform, they represent a high-use group, stakeholders with a lot to gain from deliberate, thoughtful practices of music organization.

In the specific context of research on DJs, Ahmed, Benford and Crabtree (2012) undertook an ethnographic study of DJ work to document an attachment to older media (such as vinyl) even as new forms were used. Gates, Subramanian and Gutwin (2006) analyzed DJs’ tactics during performances, including moments of improvisation to respond to changes in a given crowd’s reaction to music (see also Brewster & Broughton, 2000). Other ethnographic studies have focused on DJs perceptions of technological change (Farrugia & Swiss, 2005; Montano, 2010), noting the tensions that emerge as DJs navigate shifting technological terrain while notions of “authentic” DJ work frequently privilege adherence to older technologies. Many of these studies look at DJ work broadly, considering music collection, organization, promotion and performance. In contrast, I focus explicitly on methods of organization and music gathering that have emerged among DJs as a community of practice, a term I define in more detail below.

Theoretical Context

DJ work is recognizable as DJ work because it adheres to certain norms of performance. As Fin (somewhat cynically) explained while I was observing him perform a DJ set, “the purpose of a DJ is to provide ambience. The purpose is for the crowd to see an actual human playing music.” Undergirding that performance are much less visible but arguably more important practices of information and technological expertise. As Ron explained, “I haven’t calculated, but probably less than one percent of the job is the actual playing. Far, far less, some measly percentage of one percent is actually playing music at shows. The rest is interviewing, talking about it, e-mailing people about gigs, producing.” Although the display of a DJs skill during performance may be highly individual (indeed, many DJs stake their reputations on the original interpretations of others’ musical creations) the practices of collecting and organizing music mark a point of community between DJs in terms of shared practice, partly because they share similar obstacles in collecting digital media.

For Wenger (1998), participation in a community of practice encompasses both the "process of being active participants in the practices of social communities and constructing identities in relation to these communities" (p. 4, italics in original) where practice "includes both the explicit and the tacit. It includes what is said and what is left unsaid, what is represented and what is assumed" (p. 47). Irrespective of differences in the genre of music played, DJs constitute a community of practice through their shared experiences gathering, organizing and performing music. The labor undergirding DJ
practices is dispersed across multiple technologies, often conducted in multiple sites, sometimes alone, sometimes collaboratively. From a conceptual standpoint, this investigation of practices related to DJ work allows for critical analysis of the extent to which information practices are shared among DJs, and the relationship between technological change and DJ work.

Methods

This project takes a qualitative approach to information practices of DJs to craft a holistic understanding of how this community of practice undertakes work related to music media management. Between May of 2011 and March of 2012, I conducted 13 interviews with DJs in New York, Boston and San Francisco. Participants were located via snowball sampling, and because I was interested in comparing experiences between DJs who worked exclusively with vinyl as well as DJs who worked with digital music platforms, I screened participants to include both categories. Although I didn’t specifically seek out participants who worked with Serato, all of the participants who used an MP3 platform to perform in fact used that application. Eleven interviews were conducted face to face, one via Gchat and one via Skype. Table one provides participant details, including method of interview, location, sex and whether DJs played vinyl or Serato. This binary is tricky to maintain; all DJs reported owning and listening to vinyl, and many DJs who predominantly use Serato occasionally play vinyl or CDs during shows, sometimes playing vinyl-only sets. Rather than signifying “vinyl only” or “Serato only,” when I use the categories vinyl and Serato, it is to distinguish between DJs who exclusively play older forms of media when performing (vinyl and CDs, in the case of participants interviewed for this project) and those who use Serato, as well as other forms.

Table 1
Participant Details

<table>
<thead>
<tr>
<th>Name</th>
<th>Vinyl or Serato</th>
<th>Sex</th>
<th>Location</th>
<th>Interview method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barry</td>
<td>Serato</td>
<td>M</td>
<td>New York City</td>
<td>Face to face</td>
</tr>
<tr>
<td>Em</td>
<td>Serato</td>
<td>F</td>
<td>New York City</td>
<td>Face to face</td>
</tr>
<tr>
<td>Farrah</td>
<td>Serato</td>
<td>F</td>
<td>San Francisco</td>
<td>Face to face</td>
</tr>
<tr>
<td>Fin</td>
<td>Serato</td>
<td>M</td>
<td>New York City</td>
<td>Face to face</td>
</tr>
<tr>
<td>Jan</td>
<td>Serato</td>
<td>-</td>
<td>Boston</td>
<td>Face to face</td>
</tr>
<tr>
<td>Jarod</td>
<td>Serato</td>
<td>M</td>
<td>New York City</td>
<td>Face to face</td>
</tr>
<tr>
<td>Lola</td>
<td>Serato</td>
<td>F</td>
<td>Boston</td>
<td>Face to face</td>
</tr>
<tr>
<td>Randy</td>
<td>Serato</td>
<td>M</td>
<td>San Francisco</td>
<td>Gchat</td>
</tr>
<tr>
<td>Ron</td>
<td>Serato</td>
<td>M</td>
<td>New York City</td>
<td>Face to face</td>
</tr>
<tr>
<td>Bryant</td>
<td>Vinyl</td>
<td>M</td>
<td>New York City</td>
<td>Face to face</td>
</tr>
<tr>
<td>Carlos</td>
<td>Vinyl</td>
<td>M</td>
<td>New York City</td>
<td>Face to face</td>
</tr>
<tr>
<td>Gordon</td>
<td>Vinyl</td>
<td>M</td>
<td>Upstate New York</td>
<td>Skype</td>
</tr>
<tr>
<td>Jake</td>
<td>Vinyl</td>
<td>M</td>
<td>New York City</td>
<td>Face to face</td>
</tr>
</tbody>
</table>

Table 1 provides details on participants interviewed for this study. Pseudonyms are use to provide confidentiality. Jan preferred not to state a category for sex.

2 Briefly, Serato is a software application that allows DJs to replicate the use of turn tables with digital music files rather than vinyl records. As Fin explains: “[There is a] Serato control record, and it doesn’t just look like vinyl, it is vinyl. It has an audio signal ... so if you manipulate the record, the needle obviously transmits the same information ... Anything you do to that record, it would react exactly the same as if it were a vinyl record.” Other applications, such as Final Scratch, provide similar functionality, but participants in this study preferred Serato, believing it to be a faster and more sophisticated tool.
Interviews were semi-structured and ranged from one to two and a half hours. During interviews, participants were asked to describe how they organize their physical versus their digital collections of music, how they keep on top of new music, and how they manage changes in technology related to music and DJing. Participants who used Serato were also asked to send screenshots of the application as a documentation of their organizing structure. Using an open coding method (Strauss & Corbin, 1990), interviews were analyzed using NVIVO software. In addition to interviews, I was able to accompany several DJs to shows and observe processes of searching for and selecting music during gigs. A short paper presenting initial analysis from this project (Lingel, 2012) was sent to a small number of participants, who provided feedback as a form of member checks (Cresswell, 2007); Member checks involve asking participants to review research findings and provide feedback and are used in qualitative research as a means of validating findings (p. 217). Drawing on all these sources (interviews, participant observation, analysis of Serato screenshots, member check responses), I describe practices of organizing music as a means of tracing themes of managing digital media collections and adapting to technological change.

Findings

Practices of Organization

In talking to DJs about their music, four main categories of work emerge: locating music to play at shows; organizing music; making or producing music; and performing. In this paper, I focus on the second category, organizing music, and even more narrowly on practices of organizing digital music. In future work, I hope to address practices of locating new music in terms of everyday life information seeking, and collaboration among DJs across these categories.

It’s difficult to understate the sense of frustration that DJs feel when it comes to organizing digital music. As an illustrative example, Ron commented on frustrations of collecting global music:

It's especially bad with languages I don't understand … Like, I remember this awesome song because it has a lot of o's in it. But I don't even know how to pronounce that. And it's awesome, but I don't know how to pronounce that shit. And a lot of the stuff is in Arabic. And since it came from the Internet and was stolen, you know, I didn't get it from itunes and have it translated into English.

For DJs who specialize in playing cross-genre music, there’s a particular push to collect diverse music, but as Ron points out, this brings an additional set of challenges. Participants were often quite aware that their current practices were lacking, as indicated by Jan: “my organization is awful, it’s awful. I will name playlists based on mood sometimes, and that’s not even consistent.” Even as DJs acknowledge that their organization practices are lacking, they may have trouble altering these practices, feeling locked into the existing infrastructures that they’ve built. Obstacles to organizing digital music can be distilled into three basic categories: lack of tangibility, metadata and the size of collections, described further in table 2.

These obstacles make it difficult for DJs to organize their digital media consistently or coherently, which in turn can create problems when they’re performing. Issues of unwieldy, unstable collections are not unique to DJs; in writing about adapting systems of classification from physical to digital collections, Shirky (2005) argued that a traditional, ontological approach to organizing assets works well for small collections with well-defined categories and stable entities. They also function well when maintained by expert catalogers leveraging authoritative knowledge of the collection and serving users with well-defined needs. These attributes characterize traditional, “brick and mortar” libraries with a classification scheme (such as the Library of Congress classification system) maintained by information professionals. In contrast, Shirky argues that traditional ontologies tend to fail for large collections with loose or organic categories and unstable, unrestricted entities. In particular, such a collection struggles to meet the needs of users who are uncoordinated, heterogeneous and lack deep knowledge in the specific focus of a given collection.

What’s interesting in the context of DJ work is that as users, DJs fall into the first category of highly-knowledgeable and authoritative sources of information about their music corpora. Their

---

3 Due to space concerns, I don’t address the (il)legality of music acquisition in this paper, but it’s worth noting that one means of evading detection of copyright violations is to alter information about an asset; outside of any moral or legal concerns about consequences of downloading music illegally, there are complications for metadata and cataloging in getting music from illegal or quasi-legal sources.
Tangibility
Gordon: I can't touch [MP3s] ... I remember the way things look. I remember where things were physically in proximity to other things ... One of the things that a database cannot do is be associative ... It can't be associative in the way that humans organize things. So I feel as though I might have one record that I can associate with a number of other, different associations, multiply associated with other things.

Metadata
Ron: [Genre] is something that confounds my library skills on a daily basis, this is one of the most confusing things. I just go with my instinct ... It's an arbitrary decision. You have to have a bunch of tunes in order to justify a category, otherwise you can spread yourself really thin. Library-ness for library-ness' sake will ultimately confuse the guy that's thinking too fast and just looking for yummy bits of stuff.

Collection size
Jan: It's really frustrating, just being overwhelmed. It's just too many, too much ... When I'm going out for a club gig, I'm usually not narrowing it down to these hundred records, like I will play these hundred records tonight, only these hundred records. I don't usually do that [with Serato], so there is an element of nostalgia, for when it was that simple.

Table 2 presents three main obstacles in organizing large collections of digital music, drawing on comments from interviews. Ron makes an explicit connection between DJ work and librarianship, a theme I address later on.

collections, however, fall into the latter of Shirky's categories, typically consisting of large, heterogeneous bodies of media that often lack formal categories. Thus DJ collections may be unwieldy, particularly for people who collect across a wide range of genres, but as organizers, DJs have both the professional motivation and subject-specific expertise to develop and maintain ontological systems of classification. This points to a key tension in DJ work as related to digital organization. To consider this tension more thoroughly, I discuss the practices of the two most organized DJs I interviewed, returning to Shirky's argument throughout to frame some of the issues surrounding information management for DJs' music collections.

In this section, I focus on two participants, Barry and Jarod, as indicative of the kinds of labor that DJs are willing to invest in organizing digital music, although I occasionally bring in remarks from other participants to illustrate or complicate a point. Both participants walked me through their systems by actually showing me itunes and Serato on the laptops they use for DJing. They also each agreed to send screenshots afterwards. My point is not to generalize from these cases; in fact, these participants are exceptional in the deliberation and labor that goes into maintaining their collections. By looking at the workarounds and hacks that these DJs have created to facilitate their music, I'm interested in both identifying gaps between what DJs have and what they want, and also in looking at how the proliferation of digital media creates problems even for those with very sophisticated understandings of digital media.

Minimalism versus metadata
Barry is a fulltime DJ in New York who collects and plays a wide variety of genres, ("everything but trance") which he underscored during our interview by reading off some genres listed on his current Serato screen: cumbia, junk blues, Balkan, funk, Japanese hip hop, Brazilian, marching band. The heterogeneity of Barry's musical tastes have in part shaped the system of organization he's developed, and in which he takes a good deal of pride. During our interview, Barry seemed almost impatient for me to start asking about organization, at one point saying, "we've spent so much time talking about all this other stuff. The meat of the matter [is] I have what I think is a pretty unique system of the actual using of the music for performance ... It uniquely facilitates my unique way of music making." Barry's system hinges on two tenets – extreme minimalism in playlists and meticulous application of metadata.

Most DJs in this study utilize itunes playlists and/or Serato crates to create collections of songs they intend to play in a given night, similar to the ways in which DJs who play only vinyl will assemble crates or record bags to prepare for a show. As Em explained,

[I] pull everything up [in Serato] and then build a little crate, and it's just like pulling records. So, Serato, you know gives you a little folder and instead of me taking vinyl 12 inches and putting them in a plastic milk crate, I'm pulling MP3s and putting them in a file folder. So I'm just replicating that process.
In fact, however, building crates in Serato is not an exact replication of stacking crates. For one thing, Serato crates are by track instead of album, such that each selection of a record yields approximately 10 songs, where selecting tracks is individual. As well, many DJs struggle to create playlists that are nebulously defined (for example, by genre, mood or type of gig) and with the inconsistent application of metadata. Both in talking to DJs and watching them perform, DJs frequently flip from list to list looking for a particular track, genre, beats per minute (BPM) or mood. Frustration often arises when inconsistent metadata means that a song won’t turn up when expected. Other times a DJ will forget which list contains the track or type of track she needs. Barry’s system seeks to counter these issues, starting with eliminating the creating of multiple playlists into one list that contains everything he intends to play: “You know how people have all these huge lists of different genres? I have one … Everything goes into a big same bin and that terrifies a lot of people.”

Figure 1. This is a screenshot of Barry’s Serato library. Note the fairly consistent application of metadata throughout, particularly BPM. The inability of Serato to accommodate multiple types of genre can also be noted when Barry has had to list multiple genres in a single metadata field. The grouping category is used to denote situations in which a group of songs might be played; while most DJs I interviewed preferred to construct playlists based on this type of situation, Barry opted to keep context on the level of metadata, which he could use to narrow down his collection through sorting his collection by grouping.

Although Barry is constantly coming across new music – he spends hours every week poring over blogs and on music sites, as well as making regular visits to record stores – the list contains only those songs that he’s entirely sure he will want to play. The list does not contain songs intended for a particular gig, but rather to an overall assurance that the list contains only songs he’s vetted as playable: “I’m really picky about what goes in here ... It’s not like ‘oh, that’s kind of cool, maybe I’ll play that.’ It doesn’t go in here, ever.” Essentially, Barry pushes the problem of organization from crates to metadata, which requires him to be judicious in terms of what goes into the list for playing, as well as rigid in applying metadata that he leverages in searching the list while performing. This is perhaps one means of addressing the issue that Shirky (2005) pointed out in the problematic nature of large, heterogeneous collections – by crafting a library within a library, Barry uses his subject-specific expertise to mitigate the disadvantages of an otherwise unwieldy collection.
In particular, Barry’s system relies on BPM and genre. Metadata work begins with moving tracks into the iTunes playlist:

When I drag it into this one main [list] in iTunes, immediately, I’ll BPM it and do all the genres that I can think for it … Instead of having crates, instead of having differentiated playlists, everything is in genres. So if I have a reggae box that everything was in, I’m just fast on my fingers and type it, and so everything that would be appropriate is up when I type it … That’s pretty much it. It’s simple and people don’t get it and people don’t do it. And it works.

With metadata in place, Barry can save himself the frantic hunting through a series of playlists as well as through a series of tracks – he instead types out a search term (such as a genre) that he wants to use to limit his bin in a particular moment, and make decisions about what to play accordingly. This points back to the user side of Shirky’s (2005) ontology division – Barry is deeply knowledgeable about his own collection, and his consistent, dedicated metadata practices make him both authoritative on the music he collects and stable in the application of descriptors.

Barry is right to point out that his system is simple, in that it only requires the application of metadata. At the same time, his structurally-simple process requires a great deal of work: “I’ll go through my routine, go through all these steps, BPM, detail genres, labor over which songs I put in at the expense of taking a shower before a gig. I’m just so dedicated to it.” Based on interviews in this study, he’s also right in his assessment that other DJs don’t undertake this kind of work. Digital collections are so large that the task of metadata entry simply overwhelms most DJs. Beyond the sheer size of a collection, other problems surface from limitations of the platforms used to organize and play music in DJ work. Although Barry has a lot of confidence in his system, he nonetheless laments Serato’s current search and organization functionality:

There’s another level to take this to, and the technology’s just not there … I mean, this, I live by this thing, so it’s not inconsequential. This is how I DJ, with this box [pointing at laptop], but I need to take it to another level. I need the tagging to be more flexible, more fluid in my interface. And I need better ways to search.

These references to flexible tagging and search functionality can be leveled at both Serato and iTunes. Neither application allows for mass updates of metadata on the file level, nor are there sufficiently sophisticated searching or tagging options. Without these functions, participants develop hacks and workarounds, as evidenced by the practices of Barry and, perhaps particularly, Jarod.

Embracing the brave new digital world

Jarod is also based in New York City, typically DJing two or three shows throughout the week and hosting a weekly radio show. He has a narrower range in what he plays, often hip hop, soul and funk, but also collects across a wide range of genres. Jarod’s cataloging process is complex, beginning prior to either iTunes or Serato, with editing metadata tags of MP3s that he wants to bring into his music library. Before adding new MP3s to his main library, he brings the files into a “workspace” library, where he has set up preferences that facilitate the process of organization and cataloging:

I add the files to the [workspace] library, which causes iTunes to store a copy of each song’s metadata in its database. Then I open a separate metadata editing program called Media Rage, which I’ll use to strip the metadata out of the songs completely. This program only works on [entire] folders, which is why I need the separate iTunes library, which won’t move all of my files to their separate alphabetical locations as soon as I add them to the library. Then I use iTunes to create a different, better functioning version of the tag including the information that was just stored in its database.

As Jarod walked me through his process, I was struck by the fact that he’d set up multiple accounts on his computer for the sole purpose of enabling him to have different sets of preferences for various applications that he uses to organize metadata. Organizing music thus shapes his technological practices at a very elementary level – even when he’s not using his laptop for DJ purposes.

Where Barry relies on a single list that contains harshly-vetted materials and focuses his cataloging efforts on those, Jarod is rigorous in applying the same cataloging rules to files regardless of whether or not he intends to play them. Jarod is thorough in making sure that each asset has as

---

4 For both vinyl and Serato DJs, BPM acts as a point of convergence in linking otherwise disparate tracks. As Fin explained, combining songs based on BPM is a process in which “you’re mixing songs together, and songs match based on their tempo. Also, to a lesser degree, electronic music can be combined based on its tempo.” In other words, BPM allows DJs to play songs in sequence that are at the same tempo, and also makes it possible to layer track segments on top of each other to create new music.
complete a record as possible, utilizing all available itunes metadata fields, in some cases separating fields with delimiters in order to simulate additional fields:

I start off [the genre field] with the decade the song was released. Then, if there’s another decade that the song references, I’ll put that afterwards. So if a new song was produced to sound like it came out in the 80’s, when I search for 80’s pop, the song will come up. It’s contextual enough that it should be there even though it’s not technically from that era. Then I put all the applicable genres in order of prominence. ... I have to have the correct years for everything. Sometimes, I hate it. Say there is a compilation that came out in 2005 … I will reluctantly select all and delete 2005. Then I will be like, “ugh” and have to go look up the year for every song.

In addition to his insistence on complete and accurate cataloging records, Jarod maintains playlists full of dummy files as a wish list of tracks and records he’d like to obtain, “so if I’m searching for something, I can pull up what I have and what I want all at the same time.” For most DJs I interviewed, organizing music they already have is too much of a cataloging burden; Jarod’s process includes cataloging he’d like to have, but doesn’t actually own yet, such that for Jarod, metadata in itunes functions not only as a repository for the music that he has, but also documenting the process of acquisition.

Another practice that separates Jarod from other DJs I interviewed is his use of Apple Scripts to create smart playlists that populate based on metadata rules. Jarod first encountered Apple Scripts through consulting forums for Serato, and then realized that he could leverage scripts to create playlists that consist of tracks with metadata that don’t conform to his standards. In cataloging terms, Jarod is maintaining authority control and metadata standards by curating tracks that show up on playlists crafted through Apple Scripts. These efforts speak to the kind of expertise, stability and authoritiveness referenced by Shirky (2005) as lending themselves to ontological efforts, however, it’s worth noting that Jarod’s cataloging techniques are entirely self-enacted, shaped by improvisation and experimentation rather than institutional or professional standards.

![Figure 2](image-url)

Figure 2. This image depicts Jarod’s metadata fields in itunes. Note that the lack of available metadata fields has lead him to concatenate multiple descriptors in one field, separated with delimiters.

Although Jarod’s DIY efforts have lead to an impressive set of organizational practices, he nonetheless feels limited by the current functionality of Serato and itunes. Similar to Barry, Jarod voiced frustrations with the limitations of the technological platforms for organizing digital music. In particular, he cites the lack of flexibility and extensibility offered by itunes and Serato: “It’s all so bootleg, though,
everything I do. There’s not enough fields, so the comment field has at least seven different types of data in it and I have like a post-it on my computer that says ‘comment field tag order’ so I always do it the right way.” Nevertheless, where some DJs expressed a sense of dismay for the work of digital organization in contrast to physical assets, Jarod was notable in expressing a preference and even excitement for the opportunities of organization afforded by a digital system:

There was a point in time (a long time ago, pre-itunes) that I realized that my head couldn’t hold the details anymore. So I started marking CDs and records with markers but the limitation was always that the physical object could only exist in one location ... Also, the markings were permanent and didn’t leave room for changing my mind over time. So, when the digital DJing possibility came along, I was crazy excited about the possibilities of having my individual songs show up dynamically and being able to revise as I went. Basically when you are working with something you can only recall a small fraction of off the top of your head, your organization becomes an extension of your brain.

Most participants expressed a sense of excitement about digital media in terms of being able to gather ever larger and more diverse collections, but felt overwhelmed by tasks of organizing music. DJs are used to thinking of the music that they play as an extension of themselves (Brewster & Broughton, 2000), but it’s the ability to view organization as creative, authoritative and a necessary use of time that’s singular here. It also speaks to a series of gaps in conceptualizing DJ work: between playing and organizing music, between what applications offer and what DJs need, and between wanting access to a wide variety of music in one’s personal collection and struggling to come up with systems that enable music to be located and used.

Discussion

DJs form a community of practice most obviously in terms of live performance work. As Gordon described,

When you know DJs well enough, you don’t need to know who’s playing the records, you can just tell who’s the DJ, just by how they play the record ... These kind of imprints are the human quality of a DJ narrative, that comes with having the tools to be able to control things, and to control them in a way that affords you the ability to seamlessly put these things together. And the organization of the work is a reflection of that imagination.

Gordon’s poignant description indicates the depth and intricacy of practices embedded in DJ work. It furthermore links quite neatly to Jarrod’s statement earlier that “organization becomes an extension of your brain.” This motif of organizational practices as an extension of creativity is useful in understanding the technological stakes of DJs maintaining their libraries. I want to make two points related to issues of organization and personal media collections – first, that DJs stand to gain from thinking of themselves as a community of practice bound by librarianship, and second that systems design serving this group stands to gain from thinking of DJs as librarians.

Implications for theory: Hang the DJ, I’m a librarian

Interestingly, a number of participants used the term librarianship to refer to DJ work, or more narrowly, to refer to a specific area of DJ work. In Em’s terms: “one third of DJing is the actual being out in public, one third is knowing your technology – Serato, turntables, cables, amplifiers, speakers, all that shit – and then one third is librarianship.” For DJs in this interview pool, at least, librarianship was tied to curation of music used in DJ work. Although this work is inherently technological, both Em and Ron seemed to divest librarian work in the context of DJing from technical components:

I know a lot of DJs that are basically librarians … They’re collectors, whatever, they’re these super-knowledgeable people ... They don’t have crazy skills, don’t have a lot of knowledge about technical equipment, man, they know records, and that’s a huge part of it.

Here, Ron essentially sets up a technological divide between skills for DJing and skills of organization, somewhat downplaying the latter. This is interesting particularly given the frustration expressed by so many participants about the limitations of systems like itunes and Serato in terms of organizing music. This is in keeping with Shirky’s (2005) point that digital collections tend to be large, unwieldy and unstable, but Shirky’s solutions of opening their collections up to folksonomies or collective tagging are untenable for most DJs. Instead of thinking of themselves as librarians because of their deep knowledge
of music, DJs could start to think of themselves as librarians because they leverage knowledge of music into systems of organization that act as extensions of their creativity, which includes organizing as much as performing music. As this paper has documented, some DJs are already crafting their own organizational solutions, but far more could benefit from collective knowledge sharing about these practices. They furthermore need the technological support from applications used on a daily basis in DJ work.

**Implications for design: I’ve got the skills, now give me the tools**

Put simply, music applications intended to facilitate DJ work fail because they anticipate their systems being used by DJs rather than librarians; there seem to be design assumptions that iTunes and Serato are for playing first and organizing second, if at all. This fails to reflect some of the DJ workflows I encountered, but it also ignores the sheer vastness of DJ collections and the contingent problems of organization. DJs recognize the potential affordances of librarianship for their media collections but lack the tools to fulfill that part of DJ work. In this paper, I focused on the practices of two highly motivated DJs as indicative of the lengths to which some DJs go to build sophisticated systems of organization. My argument is not that designers should cater to users like Barry and Jarod, but to point out that more DJs would have collections as organized and structurally sophisticated as theirs if sufficient tools of cataloging and metadata were available. Applications like iTunes and Serato could start by providing extensible metadata fields and more sophisticated searching options. They should enable hierarchical crates and facilitate the process of building playlist rules. These functions reflect the needs of DJs as librarians tasked with managing large collections of digital media.

Organizational functionality could also expand beyond these applications to include other sites of DJ work. Voida et al. (2005) argued that “one of the greatest challenges for technical innovation in music sharing may be in allowing designers to make the leap between treating music sharing technologies as personal music listening utilities and treating music sharing technologies as online communities” (p. 200). The prevalent use of online music sharing platforms like Juno and Soundcloud shows the extent to which these boundaries are already quite fluid, but as is the case for most HCI research on DJ work, practices of organization are basically ignored. Sites like Juno and Soundcloud could provide platforms for explicit discussion of organization, ranging from discussions of genre to sharing organizational practices. Although willingness to share music varied across DJs I interviewed (typically out of a protectiveness of rare music), participants were far more willing to share how they organize their music. As Barry responded when I asked about talking to other DJs about his organizational process, “Sure, because it helps, it helps me be a better artist … if there’s a way for people to be better artists, it elevates the whole field.”

**References**


Searching for Theory in Metadata

Ben Li
University of Oulu
banji.li@oulu.fi

Abstract

This paper argues to include descriptions of theories alongside metadata descriptions of data. It compares several metadata infrastructure standards for interdisciplinary collaboration that include or exclude theory, and finds that metadata infrastructures for some "soft" sciences are more explicit about embedded theories than some "hard" sciences. This paper concludes by proposing to borrow an approach from social sciences for use in the hard sciences in order to explicitly address theoretical ambiguity embedded in seemingly hard data.

Keywords: metadata, theory, standards

Introduction

Two contrasting ideals of theory and empirical data are thought to advance science. Observation of low-level empirical phenomena of the specific enables one to inductively develop theories of the broad. Application of broad theories enables one to deductively predict and generate low-level empirical phenomena. In reality, both happen simultaneously and iteratively within and amongst groups of scientists. (See, for example, Morse and Mitcham, 2002; a full review of the role of theory in the philosophy of science is beyond the scope of this paper.)

The term “metadata”—data about data—implies primacy of the data-driven inductive approach building theory from the ground up, without capturing the theory used to collect and record that data. This is not a trivial point since the motivation and context for data collection must be based in some theory (somewhere between a working theory in development and an established theory being applied). A (re)user of data must understand which theories have been used to decide which data to collect and how to collect and record that data. Prevalent theories, assumptions, and practices within a scientific field may differ between the times and contexts of data collection and intended data reuse.

A primary goal of a scientific metadata record is to enable any competent scientist to examine existing (or ongoing) data collected in a foreign context, in order to judge whether the data is suitable for local (re)use (Duval et al, 2002). Metadata might describe details such as units of measurement, instruments and procedures used, and geospatial coordinates, etc. In evaluating data, a potential (re)user of data must consider objective characteristics of the data including its accuracy, precision, and other indicators of quality.

This paper argues to include descriptions of theories alongside metadata descriptions of data. First it compares several metadata infrastructure standards for interdisciplinary collaboration that include or exclude theory. Second, it highlights the roles of theory in the few standards in which theory is considered. To conclude, this paper proposes to borrow an approach from social sciences to support sustainable data reuse in other sciences. This paper's main contribution is to highlight the somewhat absent role of theory in metadata descriptions.

Acknowledgement: This work was supported by the Academy of Finland.
Copyright is held by the author.
Key concepts and definitions

The key concepts and definitions of this paper are outlined as follows, following their plain language meanings. To observe is to gather information about the state of some part of the universe according to some theory about what information is or is not interesting. To record is to create a durable inscription about the state of some part of the universe at some time. Thus, a dataset is a set of records of observations of the universe taken according to some common theory. A theory is an abstract description of the sets of expected or realized relationships in some part of the universe at some time that may be applicable to other parts or times or configurations. A (re)user is one who (re)uses a dataset collected in a previous or other context, in order to discover, develop, verify and/or apply some new theory or knowledge in accordance with scientific principles. Metadata is information that describes a set of data intended to provide context not found in the data itself; and information about containers of sets of data such as relational databases, spreadsheets, and indices. Metadata infrastructures (MIs) are the sets of human practices, tangible tools, and transistorized systems designed and used to store, process, retrieve, transmit metadata among suppliers and (potential) (re)users of data. Sustainability is the physical and social capacity to not only repeat a current practice on an ongoing basis, but to enable via that practice in support of new related practices. In this case, a sustainable MI would enable users to contribute datasets and metadata descriptions of their structures and contents, in order that a (re)user might be able to discover and evaluate such datasets for reuse in support of new theory development or applications.

While there are several interesting lines of work in the literature concerning theories of metadata (Lazinger, 2001; Greenberg, 2009) the role of theories in metadata remains comparatively unexplored.

Goals of metadata

Metadata supports a way of working in which interesting parts of the universe are distilled into discreet and unambiguous data, which can be combined with other data, in order to connect and transform formerly independent data sources into some new work of science. With very few exceptions, naked data is not professionally recognized as a first-class science output. Thus, the discovery and synthesis of data helps to enhance the qualities of such data as science outputs.

Therefore, we might compare the discovery of metadata (observed and recorded by humans and their tools at some previous point in time)—from which one could gather data to generate and/or confirm hypotheses and/or theories—to the discovery of an empirical phenomena from which one could gather data to generate and/or confirm hypotheses and/or theories. In both, we start with some pre-existing phenomena over which we have insufficient substantial control, from which we attempt to observe patterns that are meaningful according to some nascent or mature theory. Such observations must be recorded into some more concise, portable, and enduring form that acts as a discreet pointer to some phenomena (usually in the past) that is no longer directly observable.

We might then manipulate the phenomena (or our observations of the phenomena) in various ways to observe how well the outcomes of such manipulations meet with our predictions about relationships within the system of manipulated things, observed things, and recorded things.

Role and importance of theory to data (re)use

Not all environments can be observed or recorded unambiguously, so (re)users are forced to reanimate data with “metadata” (through included data standards) about how the data was observed and recorded, what the data was about, etc. For centuries, notebooks have served individual scientists by providing pointers: to one’s memories and experiences of an observation; to conditions of an observation which are salient according to some theory in use at the time; and to the outputs of observation instruments like microscopes, probes, simulation runs, or oneself through interventions at ethnographic field sites. Our reasons for observing phenomena direct our attention to some subset of observations to record and consider by some (implicit or explicit nascent or mature) scientific theory.

In as far as science is a social act—at least at the point of sharing an observation, knowledge, or discovery—a particular tree can be simultaneously: a part of an ecosystem, a member of a population of the same species, a host for other organisms, a food source for fungi, a blocker of light, a DNA sample, a climate record, an anchor in a game, a fire hazard to be cleared, an embodiment of spiritual forces, etc.
Similarly, a record of data about a tree (say, a photo, a core sample, a leaf, a height) might provide many meanings according to the theories from which it is considered and used. A social constructivist view of science would suggest that the tree and observations about it are not any of those things until we make it so for a particular extrinsic purpose.

**Challenges of theoretical ambiguity in metadata**

Was this waterflea classified as a particular subspecies because its DNA sequence indicates that it is from that subspecies, because the classification was done before other visually similar subspecies were known to science or the classifier, because that detail was found to be important to accurately and reliably observe part-way through classifying specimens in a data set, because some theory says that subspecies features would become relevant to the conclusion, or for some other reason?

Without such details about theories driving the observations, it is difficult to judge whether a set of observations could be reused for a new specific scientific purpose. This problem is compounded by increasingly more data made available online without accompanying scientific articles or context about underlying theory. Silicon-based search tools can only act on information that has been explicitly recorded, a task for which data suppliers are not generally professionally or financially rewarded. A scientist could personally contact each supplier of potential data about its circumstances of collection, but then we would question what additional value is added by an MI's extensive and elaborate fields.

The implicit goal of a good silicon-based MI, then, would be to remove or greatly reduce the human in the loop by which scientific data is discovered and shared. And for that, the MI would need to record and present something of the theories behind the data they describe.

**Comparison of metadata standards for presence of theory**

This section presents a modest comparison of (meta-)data standards intended to facilitate exchange and reuse of scientific data. The list of standards reviewed is biased toward life science fields, reflecting the author's other research interests in those fields.

**Method of comparison and rationale**

This act of observation and recording embeds several theories and assumptions, which this paragraph attempts to make explicit. Metadata systems explicitly store fields that are deemed important by the standards’ authors. This is evidenced by the prose that accompanies such a standard. Under information theory (Shannon and Weaver, 1949), absent direct communication between someone who is communicating data into an archive, and someone who is retrieving data from an archive, the only information available to the receiver is from the signal conveyed by the archive. Communications theory adds that the receiver can infer but not know for sure some additional information about the sender and their circumstances through social and cultural means. Since a major purpose of metadata systems is to enable data to be discovered and reused by senders and receivers not known to each other, the observations made for this paper examine the presence and extent of data fields in metadata standards that explicitly refer to theory behind observations (or fields into which such information could be reasonably inserted). Since both a user and an implementer of metadata tools that follow that standard must have some knowledge about that standard in order to conform to it, it is expected that key concepts and relationships of a standard are contained within the technical part of the standard (such as human-readable XML schema or formal specification) and/or in the documents that support and explain the technical standard.
Table 1
Accommodations for Theory in several scientific metadata standards

<table>
<thead>
<tr>
<th>Standard, introduction year</th>
<th>Domain</th>
<th>Theory in technical specification</th>
<th>Theory in supporting research and documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal Geographic Data Committee (FGDC), 1990</td>
<td>Geospatial data</td>
<td>Theory does not explicitly appear. Models are to be explicitly identified, for example, the geodetic model used to define locations on Earth (FGDC, 1998).</td>
<td>Supporting research and documents refer to theories of interagency cooperation and governance, causal theory and learning theory (Nedović-Budić, 2005), and various operational, economic, innovation and other theories.</td>
</tr>
<tr>
<td>Text Encoding Initiative (TEI), 1987</td>
<td>Humanities, linguistics, social sciences</td>
<td>Standard explicitly acknowledges need to be compatible with disciplinary theories (see text following).</td>
<td></td>
</tr>
</tbody>
</table>

380
Results

Table 1 presents several metadata standards showing the extent to which they address theory, ranging from not at all through detailed treatments. In most standards studied, “theory” appears neither within the standards’ technical specifications nor the supporting documentation. Likely synonyms and related terms such as “model” and “hypothesis” are also used infrequently. It is clear when those terms appear in reference to the kinds of disputable concepts scholars would refer to as theories, and when they appear in reference to (less interesting for this paper) conventions of notation. Repeatedly, “theory” referred to specific linguistic theories as in the Text Encoding Initiative (TEI), and to concepts that embed or are embedded in theories such as “goals, motivations, theory, hypotheses, strategy, statistical design, and actual work” as in EML. One would infer from the other (younger) standards’ lack of mention of “theory”, despite the clear presence of theory in the disciplines that complement those standards, that the authors of those other standards explicitly chose not to address theory in their standards.

Some metadata standards provide a field called “reference” to point to an external resource which could be cited as a theory (taxonomic identification, for example), but such external resources exist outside the standard so cannot be substantially searched related by it, can contain anything, and are inconsistent among themselves. They may or may not provide insight about their own relationships to the community. Similarly, categorization by “sub-discipline” and the like may conflict with international nomenclature for what is or is not within a discipline of a particular designation. Such references are a particular problem for data that has not been previously published since re-users of unpublished data might not be able to rely on publications to provide metadata about why and how data were collected, under what theories, etc.

Most of the above metadata standards avoid dealing with individual theories in their fields, or with theory at all. This consistent choice might itself reflect the application of some theory, but, frustratingly, standards documents themselves may only describe that a community process occurred to choose priorities, working groups, etc., not the reasoning or assumptions behind those processes. In short, a user of data described or representation in a standard could have little idea from the standards documentation alone why some observations and modes of observing are privileged above others.

However, there are two noticeable contrasts. TEI shows some explicit choices on how to handle theories to enable various (re)users of text to be able to ascertain which theories shape which data, while EML explicitly chose to acknowledge but not to theoretical diversity.

Comparing TEI and EML

In interdisciplinary work, collaborators from other disciplines might only have lay knowledge in some of the fields from which they draw data. Such re-users must be given an adequate basis from which to judge the compatibility of potential data, and the theories they embed, with their own scientific research objectives, thereby sharing responsibility for the credibility of data used.

The Text Encoding Initiative Guidelines (TEI Consortium, 2012) express several views of an underlying theory of theories used in the fields TEI supports. It is cross-disciplinary and well aware of how the disciplines interact within and among themselves. It leaves room for debates about theory, and explicitly enables observations from competing theories to be recorded on equal ground. It enables the users to specify which theories are in use in observing, recording, and interpreting the data:

“The use of the terms descriptive and interpretive about different types of encoding in the Guidelines is not intended to support any particular view on these theoretical issues,” (TEI, pg. xxvii, emphasis in original)

TEI is aware of the potential for a standard to limit work or presentation via an external theory, and it requires those (re)observing and (re)recording information to employ a similar awareness when working with both data and metadata:

“Elements such as<metaphor tenor=“...” vehicle=“...”> ... </metaphor> might well suggest themselves; but given the problems of definition involved, and the great richness of modern metaphor theory, it is clear that any such format, if predefined by these Guidelines, would have seemed objectionable to some and
excessively restrictive to many. Leaving the choice of tagging terminology to individual encoders carries with it one vital corollary, however: the encoder must be utterly explicit, in the TEI header, about the methods of tagging used and the criteria and definitions on which they rest,” (TEI, pg. 198)

The above is expressed in a lengthy document expected to be understood by all users of the standard. TEI strongly encourages suppliers of metadata to fully describe the contexts of their data, and provides obvious and explicit facilities for doing so:

“The <textDesc> element provides a full description of the situation within while a text was produced or experienced, and thus characterizes it in a way relatively independent of any a priori theory of text-types,” (TEI, pg. 468)

By contrast, EML apparently employed a different design intent, leaving the choice to provide context and theory up to users:

“The field designDescription contains general textual descriptions of research design. It can include detailed accounts of goals, motivations, theory, hypotheses, strategy, statistical design, and actual work. Literature citations may also be used to describe the research design,” (KNB, 2012)

The expectation here is also that the standard should not unintentionally constrain or promote the standard’s application to datasets that engage particular theories, but neither does it make it particularly easy for one to encode goals and motivations and theory and hypotheses and strategy and statistical design and actual work and literature to provide a thoroughly complete description, since all of that is expected to be lumped together into the same kind of generic container. That non-structured combination also hinders automated discovery of comparable studies or data that engage similar theories, methods, etc. EML goes on to further mix citations and design descriptions, such that a user must search for citations in multiple fields, and to manually distinguish among citations for research design, goals, motivations, etc. The (meta)data supplier may also be induced to introduce unhelpful redundancy by supplying the same information in both the designDescription and citation fields:

“The citation field is a citation to literature that describes elements of the research design, such as goals, motivations, theory, hypotheses, strategy, statistical design, and actual work,” (KNB, 2012)

But substantially the same information may be specified again in another way in EML:

“The researchProject complex type describes the structure for documenting the research context of a dataset or another project. It can include research goals, motivations, theory, hypotheses, etc., as well as a description of research efforts that form the basis for other work.” (KNB, 2012)

On constraints, TEI seeks to enable new discoveries to open existing constraints on data in support of new theories:

“It provides a mechanism by which the encoder can define constraints not only what it means to be a well-formed feature structure, but also valid feature structure, relative to a given theory stated in typed feature logic. These constraints may involve constraints on the range of a feature value, constraints on what features are valid within certain types of feature structures, or constraints that prevent the co-occurrence of certain feature-value pairs." (TEI, pg. 567, emphasis added)

By contrast, EML seeks to limit the kinds of observations to those already known to existing work and theories:
“The bounds element in the [bounds element, BoundsGroup, BoundsDateGroup, or DateTimeDomain] contains the minimum and maximum dates of a [numeric or dateTime] attribute. These are theoretical or permitted values (i.e., prescriptive), and not necessarily the actual minimum and maximum observed in a given data set (descriptive),” (KNB, n.d. 3)

Crucially, TEI is designed to support divergent theories within and among datasets (documents), and an open-ended "references"-like mechanism in the form of "anonymous" elements:

“… three general purposes elements … may be used to mark and categorize both a span of text and a point within one. These elements have several uses, most notably to provide elements which can be given identifiers for use when aligning or linking to parts of a document… They also provide a convenient way of extending the semantics of the TEI markup scheme in a theory-neutral manner, by providing for two neutral or ‘anonymous’ elements to which the encoder can add any meaning not supplied by other TEI defined elements.” (TEI, pg. 495)

And it does so to reconcile ambiguity with primary concern for the users of the standard:

“Texts may be described along many dimensions, according to many different taxonomies. No generally accepted consensus as to how such taxonomies should be defined has yet emerged, despite the best efforts of many corpus linguists, text linguists, sociolinguists, rhetoricians, and literary theorists over the years. Rather than attempting the task of proposing a single taxonomy of text-types (or the equally impossible one of enumerating all those which have been proposed previously), the closed set of situational parameters described above can be used in combination to supply useful distinguishing descriptive features of individual texts, without insisting on a system of discrete high level text-types.” (TEI, pg. 469)

Indeed, TEI appears sufficiently established as a coordinator that it can set its own authoritative conventions for communicating data, metadata, and theory, beyond those arising from individual stakeholder communities.

“… the terminology used in this document does not always closely follow conventional practice in formal logic, and may also diverge from practice in some linguistic applications of typed feature structures. In particular, the term ‘interpretation’ when applied to a feature structure is not an interpretation in the model-theoretic sense, but is instead a minimally informative (or equivalently, most general) extension of that feature structure that is consistent with a set of constraints declared by an FSD. In linguistic application, such a system of constraints is the principal means by while the grammar of some natural language is expressed. There is a great deal of disagreement as to what, if any, model-theoretic interpretation feature structures have in such applications, but the status of this formal kind of interpretation is not germane to the present document.” (TEI, pg. 567)

By contrast, EML is highly concerned with reconciling ambiguity with primary concern for the technological information systems that instantiate the standard:

“The information in numericDomain and in precision together constitute sufficient information to decide upon an appropriate system specific data type for representing a particular attribute. For example, an attribute with a numeric domain from 0-50,000 and a precision of 1 could be represented in the C
language using a 'long' value, but if the precision is changed to '0.5' then a 'float'
type would be needed." (KNB, n.d. 3)

Further in the TEI standard, we see topics and concerns that would be familiar to those using
EML, including information presentation which explicitly accommodates differing degrees of practical or
theoretical ambiguity:

"Even in the same field, the interests and theoretical perspectives of different
transcribers may lead them to prefer different levels of detail in the transcript
and different styles of visual display" (TEI, pg. 225)

"The distinction between <m> and <w> is provided as a convenience only; it
may not be appropriate for all linguistic theories, nor is it meaningful in all
languages. The intention is to provide a means for those cases where it is
considered helpful to distinguish lexical from sub-lexical tokens, to complement
the more general mechanism already provided by the <seg> element, …" (TEI,
pg. 535)

EML’s approach to ambiguity is to embed within the standard a hierarchy of roles who are
responsible for supplying, cleaning, and authorizing metadata and changes to it. Clearly, these two
standards are developed with different practical goals in mind about data re-use. It also appears that
these standards are developed from contrasting theories, or at least priorities, about the essential
features of a metadata standard and infrastructure for their fields.

With respect to adoption in scientific communities, TEI lists 152 registered projects and sites
worldwide (TEI, n.d.), 5,750 hits in Google Scholar, 127,000 hits in regular Google, and 13,600 hits in
popular and technical books. EML lists 297 locations of datasets in the KNB Data Catalog (not directly
comparable to TEI projects and sites), also worldwide (KNB, n.d. 2), 407 hits in Google Scholar, 10,900
hits in regular Google, and 49 hits in popular and technical books. Although TEI as a standard has a 10-
year head start on EML as a standard, both are rooted in the personal computer revolution of the 1970s
and 1980s.

The subject matter and scope of data accommodated by TEI—across humanities, linguistics, and
social sciences—would appear to be far more diverse than relatively well behaved ecological data
representable in data tables, specimen collections and annotations, images, and the like (along with aural
and textual matter from social ecology). One might expect the “soft” sciences to develop a robust
metadata system which clearly enables communications across diverse disciplines and theories. One
might even expect the ecologists to invent their own version of a well-understood computer science or
information systems tool (as they did with relational databases, and more recently, HCI). But one would
not expect the inductive and grounded soft sciences to give comparably more explicit attention to theory
than the hard deductive ecological sciences, especially since so much theory is built into the physical
instrumentation and representation of their data. Somewhat ironically, ecological researchers find it easier
to convey the essential details of their research projects and experiments in the form of a 30-second
elevator speech (the text of which would be a routine case for TEI) than to record the same details in a
16-section EML data entry tool.

And while TEI constructively handles particular theoretical debates and theory in general, as a
metadata standard it is the exception rather than the rule. Its scholarly and widespread success suggest
ways for other disciplines confronting data-sharing challenges.
Apparent core requirements/options for data about theory

In literary studies there is the concept of the “death of the author”. Not a literal death, but one in which an inscribed work may be fully interpreted without the reader or user having to know anything about the author who produced it. That is, literature is meant to be read, not written.

"A narrator should not supply interpretations of his work; otherwise he would not have written a novel, which is a machine for generating interpretations."
(Umberto Eco, postscript to The Name of the Rose)

Clearly, in science, we would want some way to express the degree to which suppliers of data and different uses of that data agree about the various interpretations and other aspects of that data, including its theories and its expressions. In particular, recording both agreements and disagreements about data, in both published works and in the metadata of the contributing sources themselves, would sustainably provide future re(users) of that data with more information and more relevant information about the offered data (this facility does not formally exist within the standards examined). As it stands, potential (re)users must independently rediscover the theories and assumptions built into data, and their (lack of) fitness or compatibility for particular purposes. Journal publications would only reveal such flaws for a small fraction of the possible combinations of offered data and new scientific inquiries. This is not a sustainable configuration overall.

A taxonomy of theories (and implicitly, a meta-theory) could certainly be developed and constantly maintained (perhaps similar to JEL classification codes from the economics literature), thereby enabling (meta-)data authors to be unambiguously explicit about the theories that go into their data (if they are actively aware of using such theories). However, that alone would not bring sustainability to the broader metadata enterprise since interactive contribution (by applying the standard to new data and metadata) would remain challenging for most potential users.

Too much of a good (meta-)thing?

The most established, if not the oldest, metadata system in science has produced metadata in abundance describing all kinds of scientific data, spanning hundreds of thousands of collections of data, each referring to hundreds to hundreds of thousands of unique packages of data and supporting information, available in no other database. Yet, according to the metadata system itself, the vast majority of this data is never reused in a way that it detects. Indeed, it is considered an excellent outcome if a package is mentioned by only three other scientists in some fields, while excellence requires dozens of reuses in others! Most of the tens (or perhaps hundreds) of millions of such thoroughly described packages of information enter repositories never to be seen by human eyes again.

This metadata system, is of course, the current system of citations within scholarly publishing. According to criteria assumed to be required for a metadata system to become successful, the scholarly citation system represents something of a best case for metadata. It has many of the conditions for which metadata champions advocate for their own strategies: broad adoption, broad coverage of all core and allied disciplines, a small number of highly versatile and time-tested standards, great freedom to adapt and localize for any research environment, well described data as verified by external [potential] users, clear chains of custody, a healthy tension between structure and innovation, etc..

One major difference between the established scholarly citation system and the recent metadata systems is the (theoretical) ongoing ability to continuously curate and revise existing data in a metadata repository, compared to the unreliable ability to occasionally add new data to the literature. The absence of universal standards to describe publications did not stop the scientific publication metadata system from succeeding. However, the resulting repository of science resources is for the most part only searchable by superficial and inconsistent appearances of concepts and theories, rather than by their contents. (Searching by a ground-up approach is great for local inductive work, but inefficient for inductive or deductive work which would benefit most from accessing broad and diverse data sources across disciplines. A robust metadata system alone is insufficient to enable bringing diverse data sets together.) Such mutual unintelligibility cannot be sustainable, as demonstrated by increasingly complex search engines that work as much by brute force as by considering metadata required to keep pace with the perpetual stream of new publications.
Another substantial difference is that the scholarly publication and citation system consists not of a single authoritative metadata infrastructure, but a patchwork of local and global systems that have been replicated and modified by users to the extent that systems like DOI have become required to describe both the published data and the descriptions of published data. Mere identical replication of metadata repositories or their contents appears insufficient to provide the same reuse value as non-identical federation of data and metadata according to several theories of how the data itself should be organized.

Extending the open ideals of metadata to provide users and non-author stakeholders with the ability to annotate and curate metadata would further leverage a potential strength of metadata to differentiate it from deficiencies of previous systems in which (difficult to scale) expert individuals and systems hold bottlenecking and gatekeeping roles.

Sustainably designing in data about theory

It is through theory that can be developed from data, and through theory that can be refined or applied through data, that both scientists and practitioners can interact with most impact. As it stands, metadata standards facilitate framing new data in support of existing theories, while largely denying users ongoing read and write access to the theories embedded in the data itself.

In addition to tracking theory, enabling annotations, and working with existing metadata systems, design theory suggests that a good MI system would reflect, facilitate, or instill desirable patterns of activity among target stakeholder populations.

But to get there, we must be able to attach value(s) to (meta)data, in order to manage it sustainably. While the philosophical position that all data is valuable tends to occur across sciences, it alone provides few insights about which data to prioritize in terms of expansion or curation in the face of limited resource and time budgets. Without a theory to reliably connect metadata to future uses ahead of time (it is unclear that such a theory would be discoverable), we must still deal with the long tail in which most of the data will be cited or reused none of the time. And we have a responsibility to minimize misinterpretation of (meta)data that we open for all to use.

What new ways of organizing thoughts does metadata require in order to be maximally beneficial? What new ways of thinking do we require to mobilize underlying large data sets? How do we fit new kinds of data or discovery into old standard metadata containers? If we want to take those kinds of questions head on, it is inadequate for metadata systems to simply bring more of the same kinds of data together. We have been doing that quite efficiently through science letters and societies for centuries, and without having to adopt additional unwieldy software systems. Metadata systems must enable new kinds of science in order to be much more than a minor sideshow. They must also enable our observations of observations (e.g., viewing of our own or others’ recordings or notes) to scale beyond a particular analysis to those spatially, temporally, and cognitively displaced from the contexts of observation.

And in light of growing scope of science, and participation in science, how do we tell the important parts of stories that go into the making of a science paper in ways that are relevant to future users of that knowledge? If scientific inquiry obtains knowledge by developing testable explanations that predict results of future experiments and observations, what is a future experiment or observation in or across metadata, and how do we explicitly support those? As taxonomic debates have shown, an ontology can suggest what might be missing within a particular view, but it cannot alone demonstrate the ways in which the view’s underlying theory might be structurally incomplete. Similarly, requiring (re)users to apply inductive or deductive theory on unstructured open field data from increasingly vast databases in order to search for reusable data might not be the most optimal arrangement of humans and computers in that particular loop.

A useful start to several of these questions would a framework or tool or theory that would permit us to concisely and explicitly characterize the theories that go into a set of observations described by metadata. Developing a narrative of all the theories that compose a discipline’s theories for observation and recording might be a good start, while a historically aware exposition the individuals’ and discipline’s tacit theories about collected and shared data. This problem does not necessarily call for a systems approach, nor for meta-theories of entire disciplines. We need only become sufficiently aware of the prevailing theories at the times of empirical observation, and later, observations of the recorded data, to state what those are in any work that draws upon re-observed data, and to be recorded in metadata systems alongside the data they describe or capture.
Conclusions

This paper has argued for the description of theory alongside descriptions of practical details concerning datasets sharable via metadata. It has compared the description of theory alongside metadata in several metadata systems. And it has outlined a requirement to explicitly describe the theories backing datasets in order to enable sustainable data reuse. Scaling up and out future combinations of human and silicon systems for interdisciplinary scientific investigation requires us to at least describe the theories that back our observations and records of ever diversifying new contexts.

References


Playfully Serious Information for Serious Play: The Integration of Community Values in an Information Resource

Julia Bullard
University of Texas at Austin
julia.a.bullard@gmail.com

Abstract

Information resources for serious leisure pursuits need to offer accuracy and usability while supporting community values such as playfulness. This paper explores how a World of Warcraft player-authored information resource integrates playfulness into guides meant for difficult game content. While written guides support play by minimizing time spent out of the game, video guides have an opportunity to reproduce the engaging qualities of gameplay, combining information and entertainment. A close reading of the Learntoraid.com’s videos and their context reveals how playfulness contributes to authenticity, credibility, and the site’s role in the community, while supporting the information seeking of players. We should expect to see the integration of community values, in this case playfulness, in information resources built within other serious leisure communities or for work and everyday life pursuits.

Keywords: serious leisure, information seeking, legitimate peripheral participation, positive information science

Introduction

A player in a competitive World of Warcraft guild is looking forward to tonight. She knows that if everything goes as scheduled, her group will reach a new boss fight, and possibly be the first on their server to defeat it. Despite her group’s previous achievements and impressive in-game capital, victory is far from assured – she’s heard that this boss can take days (if not weeks) of 4 hour sessions for a 25-player group to master. In her guild’s forums, the raid leader has linked a YouTube video posted by Learntoraid.com that they will use for the basis of their strategy, and instructed everyone to have watched it by tonight’s raid time. Headphones on, volume up, and the video fullscreen, she learns how they are going to kill a dragon.

Learntoraid.com (L2R) is a player-authored site which publishes video guides to World of Warcraft (WoW) encounters, most of which are edited and voiced by the player Kinaesthesia. L2R does not only meet the information needs of players but is in a position to shape attitudes towards the game itself, and the relationship between information and play. Using these guides is a form of legitimate peripheral participation (Lave & Wagner, 1991) in which users are cued towards the values and challenges of gameplay through the activity of navigating video information. Users are introduced to Kinaesthesia’s views on the topics of challenge and teamwork, and participate by giving their time and attention to the videos and by adding their feedback, questions, and clarification to the posts.

When players of massively multiplayer online games (MMOs) such as WoW move between the in-game world and third-party information resources, they extend the “magic circle” (Huizinga, 1955) of play into activities resembling work. Huizinga (1955) characterized play as being voluntary, superfluous, separate from material interest, yet potentially serious. In the fluid relationship between play and seriousness, a “pretend” fantasy world apart from “ordinary” life nonetheless can become gravely important to its participants (1955, p.8). In player-produced information resources such as L2R, the line is blurred between research and play.

These resources are significant sites of information and digital literacy fostered within a “leisure” community. In that these information resources represent the accumulated wisdom of a “world” of individuals, their structure, purpose, and use have parallels in information resources for the offline world.

Acknowledgements: Thanks to Dr. Melanie Feinberg and Dr. Madhavi Malapragada for their help on multiple versions of this paper. Bullard, J. (2013). Playfully serious information for serious play: The integration of community values in an information resource. iConference 2013 Proceedings (pp. 389-397). doi:10.9776/13218
Copyright is held by the author.
However, information seeking models based on user tasks (Borlund, 2003, Bystrom & Hansen, 2005, Kuhlthau, 1991) have so far been focused on work and obligation-oriented everyday life contexts, and may not explain what is taking place in a leisure-oriented information environment. The serious leisure perspective (Stebbins, 1982) describes possible motivations and rewards for engaging in the type of serious play indicative of the creation and use of these resources, and is increasingly used as a framework for information behavior (Fulton, 2009, Hartel, 2005, Lee & Trace, 2009).

The serious leisure perspective helps illustrate a scenario in which play resembles work, but motivation is intrinsic rather than obligation. Where casual leisure offers immediate rewards and requires little or no training, serious leisure is systematic, long term, and involves effort and specialized knowledge (Stebbins, 1997, 18). Several aspects of a serious leisure activity – including the need for specialized knowledge and the growth of a community – are closely tied to information seeking and sharing. In the case of information-centric pursuits, information seeking and dissemination are intrinsically rewarding and comprise the core of the leisure activity. In the case of amateur genealogy, “super information sharers” are leaders within the community (Fulton, 2009, p.764-766). These serious leisure participants, acting as information resources and mentors of information seeking for other genealogists, shape community norms of reciprocity in information sharing (Fulton, 2009, p. 765). The voluntary nature of the pursuit suggests that information management itself is an attractive pastime. In the frameworks of legitimate peripheral participation and serious leisure, information seeking is understood as part of a play activity, one that contributes to players’ understanding of the activity and their position within the community.

The core activity of endgame WoW play is a group’s coordinated execution of various complementary roles in order to defeat digital monsters, but the preparation necessary to support this heroic act includes hours of mundane gameplay and extensive research both in-game and through fan sites. Information seeking plays a major role in a group’s in-game success. By formalizing and disseminating the specialized knowledge of hardcore WoW players, resources such as L2R have an important role in dictating playstyles, shaping the information behavior of players, and shaping their attitude towards information’s role in play. In an information-centric serious leisure activity such as hardcore WoW play, it should not be surprising that information seeking can be both crucial and pleasurable. Rather, we should seek to determine how popular information resources support not only the common information needs of their community but also core values such as playfulness.

The L2R site is chosen as a case study for its popularity within the large community of WoW players and for its successful fusion of informative and entertaining content. For the community of WoW raiding guilds, this fusion is central to its status as a quality information resource. Information seeking for gameplay is itself playful; the criteria by which a strategy site is evaluated are fittingly askew from established norms of accessibility and credibility.

L2R videos do not conform to the minimalist paradigm common to much technical documentation, in which the goal is to provide quick access to material relevant to the user’s immediate task (Carroll, 1990). Instead, L2R videos combine instruction and narrative, providing information in a distinct authorial voice. In a close reading of L2R videos and their context, I will show how playfulness as a community value in the core activity extends into the supporting activity of information seeking. L2R will be read as an example of an information resource that successfully implements community values, blurring the line between information and leisure activity.

To explore the purpose and role of playable information in this environment, I will describe the challenge that strategy sites face in being serious resources for a playable game, and explore how L2R manages this dichotomy. In particular, I will show how L2R’s strengths and weaknesses in supporting information seeking are affected by the integration of playfulness. While accessibility and accuracy can be difficult to manage through video guides, the authenticity of the playable approach is positive contribution to the credibility of the information resource. In being playable, L2R videos blur the line between information seeking and gameplay, and are in a position to teach players about more than just in-game mechanics.

L2R

L2R was launched in December 2010 and was updated with new video and textual content several times a month. When last viewed in July 2012, the site and its YouTube channel contained 99 videos, most of which are strategy guides to scripted battles in WoW, with the remainder comprised of editorial content and guides to specific game modifications. L2R and its main competitor for WoW video guides, Tankspot.com, have similar view counts on each of their YouTube guides. The L2R channel has
more videos over the same period due to covering many versions of each fight, including the various in-game settings for difficulty (Normal, Heroic) and number of players (10, 25).

The site also includes environmental monitoring content such as an embedded livestream of a contributor’s computer screen, guild rankings generated by another site, links to recent podcasts from a contributor, and server status for the game. A version of the site from February 2011 can be viewed through the Internet Archive’s Wayback Machine, and is nearly identical in layout and design to the version discussed here.

The L2R site is structured as a blog, with a title and menu header and a right-hand column with links, and blog entries organized from newest to oldest in the main column. Videos are hosted on the L2R Youtube channel and embedded in the blog posts. Blog entries include the embedded video, along with a written outline of the encounter and a transcript of a section of the video. In the main page of the site, the blog entries are cropped to show the embedded video, a brief summary, the L2R difficulty ranking, links to the full blog post, the forum thread related to the post, and the categories assigned to the post.

The strategy videos display recorded in-game footage altered to include instructional text and graphics, one or more songs as background, and voice-over commentary and instruction by the player Kinaesthesia. Kinaesthesia is the identified author for all strategy guides for raid encounters, the most difficult tier of content in WoW. The videos are between 5 and 30 minutes in length, depending on the length and complexity of the encounter described. In the videos, Kinaesthesia introduces the encounter, describes the abilities and phases that comprise the encounter, and then describes a strategy for successfully completing the encounter.

The site is organized to provide effective access points to the videos for raid content, even at the expense of navigating among the editorial content and guides to less difficult types of encounters, indicating to users the relative value placed on raid encounters. Because the L2R site has multiple authors and is produced in a blog format, the primary mode of organization is adding category tags to posts, rather than creating a hierarchical menu of posts. Kinaesthesia labels his video posts within the site according to the series of encounters it belongs to, as well as the categories “Raids”, “Raiding”, “Videos”, “Raid Guides”, and “Guides”. No other type of post on the site has so many access points. Raiding content is privileged while all other posts are collocated as a mess of “Other”.

### Information for Serious Players

Raid encounters are exclusive to player groups who have the preparation, coordination, time, and membership to attempt the 10- and 25-player fights. Raid encounters are the most ambitious and demanding scripted battles in the game, both for the in-game materials that must be accumulated to attempt them and the complexity of their events. Appropriately, they also confer the greatest rewards. Hardcore players – who orient themselves towards competitive achievement in this game content – may be reviled in some player communities for hours, days, and even weeks at a time. Information needs in this context are often essential to the core activity, sometimes urgent, and always part of a series of social relationships. Hardcore guilds do not approach encounters with trial-and-error but with strategies derived from personal experience and community research. More accurate information can sometimes be the difference between a coveted “World 1st” achievement and obscurity, and no guild can compete without every member of the group being well-informed in advance of their role in the overall strategy. By virtue of his early access to these encounters, the clarity of his explanations, and the effort he puts into video production, Kinaesthesia is seen as a credible source that can effectively guide individuals and groups through these situations.


**Written vs. Video Guides**

Despite these grave stakes, WoW is a playful game, distinct from the real world in tone and aesthetics. The visuals are cartoonish, enemy character dialogue is often absurd, and game content is riddled with puns and hidden jokes. Strategy guides need to support this playfulness while providing serious instruction. Usability, accuracy, and credibility are necessary for the effective dissemination of game strategy, but the information is meant to serve play in the recreation space of a fantasy role-playing game. Strategy guides in WoW and other MMOs are usually presented as written guides (as in the case of wowwiki.com and icy-veins.com) or as videos (as in the case of L2R and tanksot.com), and each supports playfulness in different ways. A site such as Icy-Veins manages the serious-play conflict by providing clear and to-the-point written guides that facilitate quick use and bring the player back to the game with minimal interruption. This approach privileges play by removing barriers to play – it provides information and strategy as clearly as possible, providing all relevant details for the player to synthesize with their experience of the fight.

Written strategy guides, unlike video guides, can be scanned and searched in order to focus on a particular detail relevant to the user. Written guides do not have to be read in the linear form in which they are presented, especially after the first read-through when the user is likely to want to confirm specific details as they encounter them in-game. In fact, written strategy guides can be easily consulted during game play, by printing them out or by quickly switching between the game application and a web browser. Additionally, players can review written guides at times they cannot play the game, such as when they do not have broadband internet access, when they cannot use audio features, or when reading text on the internet is more acceptable or less conspicuous than gameplay, such as in a school or work setting.

Videos sacrifice the usability of searching, scanning, and accessibility for entertainment and engagement. They privilege a linear viewing, cannot be searched, and require significantly more effort on the part of the author to update them when new information is available or changes are made to game content. It would seem that, based on the gravity and urgency of the information needs of this player community, written guides would be the dominant mode of information dissemination. However, this is also a community of play, and its information resources cannot be assessed based on accessibility and usability alone. Playfulness, as a community value, must also be considered. While both written and video guides can integrate playfulness into information dissemination – such as through authorial voice and the visual aesthetics of their website – video guides have an advantage in expressing some of the fun of the game itself. These videos are an enjoyable experience in their own right and authentic to the activity they support. By examining the style of L2R videos and user commentary, I will demonstrate the role of playfulness in a serious leisure information resource. Video guides can incorporate some of the usability of written guides, and the resulting product elevates both play and information by blurring the line between leisure and information seeking.

**Recreation & Information**

The primary conflict managed in gaming strategy guides is between knowledge and execution – the importance of knowing about the game against the skill to press buttons correctly under pressure. As an information resource, L2R implicitly values the role of information in gameplay. The effort put into creating engaging and detailed strategy guides indicates the importance the author places on disseminating his strategies to the player community. However, the instruction provided is to assist in the successful performance of a correct strategy, which puts knowledge in the service of execution. Kinaesthesia’s narration describes a particular relationship between knowledge and execution, one that foregrounds the importance of focused performance by naturalizing the value of understanding and preparation. In these videos, the synthesis of information and playfulness elevates both.

L2R privileges play by supporting gameplay and by being in itself playful. Kinaesthesia’s strategy guides require more channels of attention from a user, who must view the video – at least the first time – in a linear fashion. This demanded attention is not necessarily an obligation or burden; though many users have a clear information need that brings them to L2R, the production and humor of the videos makes the immediate experience rewarding, as well as promising future enjoyment in gameplay. This is not to say that the guides are not serious – they are. The users value the videos for their accuracy, and place enough trust in Kinaesthesia to implement his strategies in their own play.
It is not miraculous that the play and seriousness of L2R do not interfere with each other; rather they are parallel functions that serve each other. For example, Kinaesthesia’s convention of opening each video with a pun does not just soften the blow of listening to 8 minutes of strategy. The puns are not random but function as mnemonic strategies for boss encounters. Kinaesthesia’s jokes exploit the most iconic or essential mechanic of each fight, providing a shorthand around which a user can organize their growing understanding of the encounter. For example, the video for the boss Lord Rhyolith presents a pun on “soul” and “sole”, a mnemonic device for an event that revolves around moving a monster’s feet. Within what seems like a dreadful pun, Kinaesthesia cues the player to the major mechanic of the fight.

Several of the Top Comments in the YouTube pages for L2R videos express sentiments like “I haven’t done this fight on normal in months but I still watched the guide for puns” (Beachbum1546 in response to LearntoRaid’s Baleroc Strategy Guide (25 Normal)). The users emphasize their appreciation for the performative aspects of the guides, including the production quality, jokes, and music, by stating that the information content is irrelevant to them. Video viewers see these Top Comments “above the fold” and are therefore aware of this alternate use of the resource, whether or not they came to it in a directed act of information seeking. It is possible that this awareness shapes their interpretation of the value and purpose of the strategy guides, and cues new users to treat the videos as something more than information.

These comments are a strong example of how an information resource can transcend its informativeness into playfulness. The constant jokes and attitude in the videos do not just make the information palatable, as some users describe, but make the video enjoyable in its own right. This is a tangible contradiction of the enduring belief in information science that information seeking is always problem-driven (Allen, 1996). In the case of users who value the videos regardless of the pertinence of the information presented, the information resource does not solve an information problem but first entertains.

Case explored this blurry line between information and entertainment (2002, p. 108-115), demonstrating how some forms of media can perform different functions for different users, or both to many. Whether it is news being sought for entertainment, and fiction imparting factual knowledge, “the best of both worlds” occurs when “hard facts are presented in an entertaining manner” (p. 115). Strategy guides for video games are an excellent example of this duality – they are produced in anticipation of a critical information need, but in the context of a leisure activity. User comments in response to L2R videos show that the resulting product is both entertainment and information. The sacrifices of usability to engagement described above are reshaped in this scenario: searching or scanning is not “missing” from the videos, because skipping sections would detract from the complete experience in which information about play becomes entertainment. For the users who have an information need related to the strategy guide, the video can become an enjoyable means of study, or one that frustrates them for its shortcomings in usability. However, as the next section will describe, some of the usability of the written guide can be recouped, producing an information resource that better serves the information seeker without destroying the narrative flow of the video.

**Recouped Usability**

Kinaesthesia has sacrificed the usability of written guides in favor of the spectacle and engagement of the video form, but some of these sacrifices are recouped by taking advantage of YouTube functionality and providing transcripts. The blog entries in which the videos are embedded contain a detailed written guide to the encounter, with the first half presenting the abilities and events in point form, and the rest a word-for-word transcript of the strategy section of the video. Whether the user accesses the L2R video from the site or from YouTube, they will have access to a written version that can be searched, copied, and read with the conveniences already attributed to written guides.

Additionally, Kinaesthesia has increasingly taken advantage of functionality introduced by YouTube. Beginning on February 15, 2011 with the Maloriak video guide, Kinaesthesia began including timestamp links in the information attached to the YouTube video. This feature, introduced in 2008 and improved in 2010, allows users to create links that play a YouTube video from the minute/second designated. Though the base functionality existed before L2R’s YouTube channel was created, Kinaesthesia did not include timestamp links in YouTube postings until a few months after the improved functionality was introduced. The timestamps divide the video into sections – such as “Abilities: Phase 1” and “Outro” – that allow the user to navigate to the section of the video that meets their information need.
Along with the transcript, this creates a crude but practical adaptation of the search/scan function native to written guides. The video form of the strategy guide will likely continue to become more accessible as platforms such as YouTube introduce more navigation and notation features.

L2R users also facilitate the usability of video guides, particularly by improving the accuracy of the guides by correcting Kinaesthesia or emphasizing details not covered but that they consider valuable to success. That these comments are often voted to the Top Comments section of the page suggests that this interaction is valued, and accuracy of information is important to Kinaesthesia’s audience. This value for accuracy is reflected in the few “Uploader Comments” that appear among the L2R videos. In a handful of cases, Kinaesthesia responds to a specific user or pre-emptively corrects information based on changes in game mechanics. While it is time consuming for Kinaesthesia to update a video with new or corrected information, responding to the video with a written correction is a reasonable compromise. The information in the video is still static in comparison to an editable written guide, but the affordances of the YouTube page allow for ongoing discussion.

**Authentic Playfulness**

Finding, watching, and commenting on these videos are among the first tasks a player takes on in becoming a raider. As a form of legitimate peripheral participation (Lave & Wenger, 1991), information seeking introduces new players to expert players’ concerns, the language they use, and their approach to play. L2R is in the position to set player’s expectations on what constitutes raiding, and its playful approach highlights the “epicness” and intensity of the experience. Video guides accomplish this transfer by closely resembling gameplay itself, in that their primary visual content is recorded in-game encounters. They are more engaging, and more immersive, in that they require the user’s full attention to glean all the information needed. The user needs to listen to the voice-over commentary, read the instructions as they appear, and relate these to what is happening in the gameplay displayed. Videos are more game-like than instruction-like: they visually resemble the game and demand similar modes of attention from the user. This sense of the “game-like” is enhanced by the music added to the background of the verbal commentary from Kinaesthesia. The genre of music chosen – usually rap or dance music – adds a sense of bravado and urgency to the presentation, allowing for a vicarious experience of the excitement of gameplay.

L2R videos recreate the intensity and complexity of the raiding experience by drawing attention to the wealth of information involved in an encounter. The recorded gameplay used as source material for the visual content of the videos gets constrained and overlaid by the progression of textual information throughout the video. The bottom portion of the user interface, usually reserved for the player’s onscreen mapping of keyboard inputs, their text chat, and map, is replaced with a large black bar upon which bullet-point information is given. The L2R watermark remains onscreen throughout the video, and emphasized textual information is briefly superimposed over the remaining gameplay. The gameplay itself is not the linear gameplay of a kill video – in which a group proves a successful encounter through recorded gameplay – but looped sections of video chosen for their illustration of specific mechanics and strategies. Play, as represented by the gameplay video, is literally background to and superimposed by information. To non-users, the finished product might seem “busy”, but raiding requires managing varied visual and auditory input, keeping track of changing cues throughout the screen, and even adapting user interfaces to maximize situational awareness. In this way, the accumulation of verbal instruction through the L2R video is seamlessly game-like, as it demands the same modes and intensity of attention that players experience during the most satisfying game encounters.

By being a staple of raid preparation, L2R videos are in the position to define raiding itself. For example, Kinaesthesia manages the expectations of the players by characterizing the most difficult encounters as punishing or arduous even in the opening joke. In these strategy guides, Kinaesthesia instructs players not to expect an easy victory, but to prepare themselves and their raid for weeks of attempts as they learn the encounter, and instructs them on how to adjust their goals and scheduling to avoid frustration and burn-out. This can be seen as meta-strategy, as it provides a strategy for managing morale and confidence in order to be effective in-game and remain a cohesive group despite continued disappointment. In these moments, L2R does not just provide information that improves individual gameplay, but supports the social cohesion of player groups in the community. Though a player might approach this source to learn their role within an encounter, they also learn the practices of good leadership and teamwork. From outside the game itself, a resource such as L2R can define the
boundaries of play: in narratives such as these, both defeating digital monsters and managing a group of players is “the game”.

In one remarkable case, a video guide was published before Kinaesthesia’s group had successfully completed the encounter. In the forum thread for the Ragnaros 10-person strategy guide, another member of Kinaesthesia’s guild states that the video was created and uploaded once they had seen all of the fights mechanics, but before they – or any other guild – had defeated it. This is not an admission that endangers the credibility of the video, but one that reinforces L2R’s function as a community service rather than an expression of personal success. Whereas “kill videos” are celebrations of mastery and proof of success, the L2R videos have a collaborative rather than competitive function within the community. By teaching other players how to raid, L2R contributors risk training their replacements, but this is consistent with a serious leisure career in which sharing information is a component of expert status (Fulton, 2009, Lee & Trace, 2009). In this particular case, the risk involved in disseminating this knowledge is tangible, which demonstrates to new members that sharing expert knowledge is also a community value. In the contexts of many games, using information gleaned from another player’s experience would be called cheating, even by those that condone it (Consalvo, 2007). L2R and similar resources make clear that at least the dissemination of such information is positive behavior for this community.

**Playful Credibility**

These playful qualities contribute to the authenticity of L2R videos and to Kinaesthesia’s credibility. The playfulness that characterizes the L2R guides does not appear to be detrimental to Kinaesthesia’s credibility, because his authority is not based on maturity or solemnity. Whereas surface features such as a professional tone might be important in other information resources, it is the opposite that makes the dense information in these videos palatable. Even his association with the guild “vodka”, a well-known guild that competes among the most accomplished, is not invoked on the main page of the website or the YouTube channel. Kinaesthesia’s credibility is instead established by the quality of his video editing and the presence of gameplay footage in the videos. Experienced WoW players can evaluate the information quality of the video — recognizing it as “useful, good, current, and accurate” — rather than relying on an evaluation of information credibility — the trustworthiness of the message source (Savolainen, 2011). Authority in this case originates in the information itself.

The source and style of the information contributes to the status of its author. The production quality of the L2R videos, which improves slowly throughout the collection, indicates the effort put into the videos and is a tangible manifestation of a skillset that distinguishes Kinaesthesia from the general player population. The source material for these videos and their visual polish helps establish Kinaesthesia’s position as an expert member of the raiding community. Users that describes the visual polish of the videos as “just incredible” (YouTube user superpudd1 in response to LearntoRaid’s Ragnaros Strategy Guide (25 Heroic)), express a respect or even awe of Kinaesthesia that could influence their judgment of his knowledge of the game. The visual source material Kinaesthesia uses as the background of his video reinforces this trust. By virtue of having had access to these encounters, Kinaesthesia is already among a select subsection of the player community who have the experience and expertise to reach this content, let alone manage it successfully. In fact, the L2R videos rarely if ever show the conclusion of the encounter. Kinaesthesia could provide further proof of his authority to dictate strategy by including the final, triumphant moment of recorded gameplay, but this does not contribute to the function of the guides and is not included. Rather, Kinaesthesia closes each video by prompting the viewer’s own attempt, positioning the video itself as merely a prologue or warm-up to the main event.

**Conclusion**

I have characterized the production, design, and organization of strategy guides on the LearntoRaid site as a success of playful information. There are modes of play that are diminished in Kinaesthesia’s focus on raid content, and to casual players, this information might seem anything but playful. However, among Kinaesthesia’s audience, there seems to be an appreciation for both the accuracy of the information and the playful and artistic fashion in which it is presented. By examining an information phenomenon that transcends problem-solving, negative affect, and problem-driven
information needs, I have provided a tangible example of positive information science (Hartel, Kari, Stebbins, & Bates, 2009). In the form of video strategy guides, information produced through play supports information seeking for gameplay but also becomes an entertainment source in its own right. It also teaches players how to value their gameplay, how to approach challenge, and how to manage frustration among a team. I have described an information resource for a gaming leisure community, but the variety of purpose found can and should be possible in information environments for work and everyday life. We should expect to see community values integrated into community information resources – not only in content but in style and tone. For a leisure community, this is particularly tangible, as there are identifiable qualities that draw members to a voluntary core activity. However, this should not be exclusive to leisure resources; rather, this integration of community values will also be the case in the most engaging work and everyday life information resources. Information systems always are always value-laden (Bowker & Star, 1999, Feinberg, 2007, Mai, 2010), and information professionals have opportunities to respect and express community values through a variety of design decisions. We can also reach better understandings of communities by examining not only what information members seek and use, but how they organize and present this information for their peers. Particularly through exploration of engaging and enjoyable information phenomena, information science can grow to understand how information can be a positive aspect of our lives and design information resources with this goal – not just the solving of problems – in mind.

References


"The Public is Part of the Audience": Information Access for Citizens, Democratic Accountability, and Climate Change Science

Shannon M. Oltmann
School of Library & Information Science
University of Kentucky
shannon.oltmann@uky.edu

Abstract

Information access has been identified by several scholars as an important research area within information science; with intellectual, physical, and social components, genuine access to information can be difficult to secure. According to the theory of democratic accountability, information access is particularly important for citizens in a democracy, who must have access to public policy information and the information that informs policy decisions in order to hold their elected officials accountable. This research conducted two case studies into the accessibility of climate change research utilizing qualitative semi-structured interviews. Respondents indicated that citizens' intellectual, physical, and social access to the information in question was curtailed, thereby hindering their ability to practice democratic accountability. This research extends previous work done in information access and scientific research and demonstrates a need for further investigation into citizens' access to government and scientific information.

Keywords: information access, scientific communication, qualitative, democratic accountability

Introduction

According to some scholars, the right of access to information has become the dominant right in the information and knowledge era (Lor & Britz, 2007, p. 392; see also Byrne, 2003; Doctor, 1991, 1992; Jaeger, 2007; Lievrouw & Farb, 2002). Over 20 years ago, Buckland (1991) recognized that access emerges as a recurrent theme across information science research (p. 77). This research adopts Jaeger and Burnett’s (2005) definition of access as the presence of a robust system through which information is made available to citizens and others (p. 465). Here, system entails the socially and politically contextualized complex means by which individuals obtain information (Burnett & Jaeger, 2008; Burnett, Jaeger & Thompson, 2008; Jaeger, 2007; Jaeger & Bowman, 2005). Hence, this definition reaches beyond technological tools and is useful for studying various forms of information access. Such a system has physical, intellectual, and social components.

Lor and Britz (2007) made a similar argument: a well-developed and well-maintained information infrastructure alone is not enough. The information that is accessible should also be affordable, available, timely, relevant, readily assimilated, and in languages and contexts users can relate to and understand (p. 390). The relationships among these different components of information access were explicitly problematized when Blakemore and Craglia (2006) wondered whether physical access to information is being prioritized above that of social access (p. 19). Thus, other scholars have recognized, more or less explicitly, the physical, social, and intellectual components of information access.

Burnett, Jaeger, and Thompson (2008) explained that physical access includes the physical structures that contain information, the electronic structures that contain information, and the paths that are traveled to get to information (p. 57). Geography, technology, and economics can all affect physical

Acknowledgement: The author would like to thank Stacy Kowalczyk and Howard Rosenbaum, as well as the anonymous reviewers, for their suggestions and critiques.

Oltmann, S. M. (2013). The public is part of the audience”: Information access for citizens, democratic accountability, and climate change science. iConference 2013 Proceedings (pp. 398-408). doi:10.9776/13221

Copyright is held by the author.
access. Intellectual access refers to understanding information in a document, including how the information is categorized, organized, displayed, and represented (Jaeger & Bowman, 2005, p. 67). Individual traits such as physical or cognitive abilities and disabilities, language competence, and technological literacy can affect intellectual access; whereas physical access is enhanced, constrained, or manipulated in the external environment, intellectual access is affected by the individual's internal characteristics. Finally, the concept of social access suggests that elements of one's social world, including social norms and worldviews, influence the information one accesses, and how and why particular information is sought (Burnett & Jaeger, 2008; Jaeger & Thompson, 2004). Normative social behaviors, according to Burnett, Jaeger, and Thompson (2008), define the appropriate mechanisms and activities involved in information access (p. 59).

Jaeger (2007) utilized case studies to illustrate how political contexts can affect information access. This research expands upon this stream, using case studies from climate change science to investigate how various political and social considerations affect access to information, particularly scientific research.

**Theoretical Background**

In a democracy, sovereignty rests with citizens, who delegate decision-making authority to elected representatives in a principal-agent relationship (Behn, 2001; Bovens, 2006). Thus, Heritier explained, in the context of the main form of representative/republican democratic legitimation, transparency and access to information play a straightforward supportive role. They function as a prerequisite for exercising popular control over government activities (p. 824). In other words, information, transparency, and accountability are all central to ensuring the legitimacy of democratic government. Transparency is the conduct of public affairs in the open or otherwise subject to public scrutiny (Grimmelikjuijisen, 2009, p. 175). Soderman (1998) added, as far as possible, the information on which the decisions are based should be available to the public. The theory of democratic accountability posits that citizens must have access to policy decisions and the information that informs such decisions in order to hold their elected officials accountable.¹

In our modern democracy, scientific research is often the basis for such policy decisions. Frequently, scientific research may be intricately linked to public policy; scholars in this area recognize that most public policies and political decisions depend at least in part on some scientific or technical analysis (Kelly et al., 2004, p. 5). There is growing acknowledgment that science may clarify political choices, but [it] can never depoliticize them. Empirical studies can identify the tradeoffs, but they cannot evaluate the choices (Smith, 2005, p. 295; see also Brosnan, 2007; Doremus, 1997, 2008; Jasanoff, 1990; Kelly et al., 2004).

Ensuring that citizens have access to the scientific research underpinning policy decisions enables citizens to be better informed about their government and to better hold officials accountable. Kulawik (2009) argued, although accountability's significance for a democratic polity on the whole seems obvious, applying the concept to science and science policy is quite a recent phenomenon (p. 470). Thus, this research utilizes the theoretical lens of democratic accountability to query citizens access to scientific research. Citizens need physical, intellectual, and social access to the science that informs policy decisions in order to evaluate the actions of their elected officials and hold them accountable.

**Methodology**

Two case studies were conducted to investigate the ways in which political contexts affect the public's access to scientific information. The process of case identification and selection began with the Union of Concerned Scientists (UCS), which has collected reports of incidents in which access to or use of scientific research was blocked in some way. Two of these incidents lent themselves to a case study of information access: they were both about climate change, both contained names of several key individuals, and both contained explicit restriction of access to scientific information.

Key individuals from these cases were initially identified from the Union of Concerned Scientists website (2010) and media reports. Additional respondents were located through relevant documents and

¹ Generally, this theory does not consider whether citizens do actually use said information to hold elected officials accountable.
snowball sampling. Semi-structured interviews were then conducted via telephone. Interviews are an effective method for probing beneath the surface, soliciting detail and providing a holistic understanding of the interviewee’s point of view (Patton, 1987, p. 108). While there has been some concern that telephone interviewing might produce less data, the researchers who have compared telephone interviewing with field [face-to-face] interviewing have generally concluded that telephone interviewing was an acceptable and valuable method of data collection and was successful in obtaining completed interviews. (Sturges & Hanrahan, 2004, p. 110; see also Cook, 2009; Holt, 2010; Novick, 2008).

The semi-structured interviews were conducted with a framework of questions which asked respondents to describe the incidents in question, discuss their roles, and consider how information access was affected. Questions were left open-ended; each interview built upon a particular respondent’s background, knowledge, and responses. Further probing questions were used to deepen the response to a question, to increase the richness of the data being obtained, and to give cues to the interviewee about the level of response that is desired (Patton, p. 125). Meho (2006) notes that with qualitative interviews, researchers seek to improve understanding of social and cultural phenomena and processes, rather than to produce objective facts about reality and make generalizations to given populations (p. 1284; emphasis in original). Thus, the interviews yielded in-depth consideration of the research area as seen through the respondents’ perspectives.

All of the principal actors in these cases, a total of eleven people, were contacted, and eight agreed to be interviewed.2 The eight interviews, across two cases, lasted a total of 5 hours and 38 minutes. With the respondents’ permission all interviews were audio-recorded, transcribed, and then uploaded into QSR NVivo 8, software which allows for iterative coding of qualitative data. Among other features, NVivo allows researchers to group similar items, create hierarchical trees of nodes or features, and code directly onto the content. The data was initially coded into broad categories.3 As common themes emerged from iterative reading of the data, new categories were created. The coding categories were then verified by an independent reviewer (Kohen’s kappa .83) who sorted a random sample of data into established categories.

In addition to the interviews, one case was supplemented by a publicly available Congressional deposition. Philip Cooney was the Chief of Staff for the White House Council on Environmental Quality. When contacted, he said I cannot comment publicly I wish I could help but I will not. He then mentioned his deposition before the Congressional Committee on Oversight and Government Reform; this publicly available information was added to the relevant case. In order to protect the anonymity of the other respondents, none of the public statements they made were used.4 Because the respondents’ own words are the actual data (Patton, 1987), the analysis relies heavily on the interviewee’s speech. Their names have been changed to protect their identities; similarly, their job titles have been obfuscated somewhat (and, as mentioned previously, no publicly available testimony from respondents other than Philip was used).

Case #1: Climate Change Reports Distorted

Here, key findings from the two cases are presented, prior to synthesizing the analysis of both cases.

This case involves two different types of editing actions5 made to several climate change related documents. First, many reports about climate change science were edited to increase uncertainty or downplay the harmful implications of climate change. Second, references to the National Assessment of

---

2 Because the cases involved a relatively discrete event, the possible number of respondents was small. As with all interview-based research, the availability of willing respondents affects the breadth and depth of the research that can be conducted. With the type of politically sensitive topic under investigation here, this research is more dependent than most on willing (and courageous) participants.

3 The research presented here was part of a larger research project, so not all of the coding categories are discussed in this paper.

4 Many documents pertaining to these cases were studied and analyzed; overall, the textual data did not significantly add to the cases certainly not enough to risk respondents confidentiality. In the interests of protecting respondents, some of whom are still federally employed, texts were not used in the case analyses.

5 For examples of the edits in question, see http://www.nytimes.com/imagepages/2005/06/07/politics/20050608_climategraph.html.
Climate Change (NACC) document, a seminal, broad report, were removed from almost all subsequent reports.

Four individuals involved in the controversy agreed to be interviewed. Lionel was a career bureaucrat, focusing on policy aspects of climate change science. His frustration with this situation led him to resign his position and publicize the disputed editing. Kevin was an alleged censor, a political appointee and senior official with decades of experience with climate science, both in and out of the federal government. Regis, similarly, was cast as an alleged censor, a political appointee and senior official at the White House’s Council on Environmental Quality. Joan had a minor role, as a non-governmental observer and climate change professional. In addition to these respondents, the Congressional deposition of Philip was used because his role was significant. From the text and interviews with these respondents, four areas of disagreement emerged: how to classify the content of the document, the level of uncertainty about the science, the significance of editorial changes, and the role of politics.

**The type of document.** The climate change reports that received the controversial editing included the 2003 draft of Our Changing Planet, the annual report to Congress, and the 2003 Strategic Plan for Research. These documents were variously described by the principal respondents as a major effort to connect scientists and other experts and stakeholders on diagnosing the implications of climate change and impacts on the United States (Lionel), as policy documents that were seeking to describe the scientific basis for, sort of, the next steps of policy development (Regis), and as a very broad assessment plan which tried to define the questions that needed to be answered.

The Strategic Plan was created by Lionel and his scientific staff; they pulled scientific analysis and conclusions from several different sources, primarily articles published in premier, peer-reviewed scientific journals. For Regis, this path to the Strategic Plan is important; he saw the document as a policy document, not a scientific work. He explained that the irony is that the debate and the fascination was over a policy document and its efforts to summarize science...The science occurred, the science was funded, peer reviewed papers were produced and yet they called it manipulation of the science. In contrast, Lionel saw the editing of the Strategic Plan as a direct assault on science because the document, in his view, contained solid statements about climate change and the human influence.

**Scientific uncertainty/factual basis.** Respondents also disagreed about the levels of uncertainty regarding the science. Kevin said the document and the science it summarized appeared to be relatively unsettled. In the Strategic Plan, he hoped that the group could begin to address the questions in terms of what do we know, what more information do we need and the like (see also Philip, p. 73 of his testimony). However, Lionel, the alleged victim, saw the document as a way to communicate the facts of climate change to a broad cast of stakeholders. Once the impacts of climate change are translated into real things happening to real people in real places, then that can generate more public pressure for stronger government response (Lionel). Both Joan and Lionel believed the scientific evidence of human impacts on climate change was irrefutable, convincing, and urgently needed.

**Significance of editorial changes.** For Lionel, any attempt to change those solid statements was an attempt to manipulate the scientific basis of those statements. Lionel called these editorial changes the political policing of the research program. He reported that the suppression of the NACC, which included no reference to it allowed in program documents, no discussion by program officials, no use of the findings in research planning, abandonment of the scientist-stakeholder network, and no follow-on work as the most harmful interference with climate change science. Lionel viewed the National Assessment as the definitive statement on climate change science and policy and thought it should serve as a foundation for subsequent reporting and activities. From Lionel’s perspective, the editing was so severe that he resigned; he said he finally realized that if you’re really interested in honest climate change communication, your hands are really tied in that office. So I resigned. He distinctly saw purposeful, malicious intent behind the editorial changes to the Strategic Plan and other reports.

Kevin sharply disagreed with this portrayal; instead, he argued that the scientists were heavy-handed and strong editing was needed to keep the Strategic Plan a neutral, effective policy tool. He said, Many of what I might call the activist scientist group viewed that they couldn’t let any paragraph or sometimes almost any sentence pass without making it what I called an advocacy sentence or paragraph. He argued for the removal of adverbs and adjectives from the reports. Kevin believed that the climate change scientists, as a result of working on this problem for years, were no longer neutral.
Regis, the most senior of the three, saw the Strategic Plan editing more simply than either Lionel or Kevin. Regis portrayed the editing as a standard procedure often done when multiple agencies are contributing to a report. Philip agreed, noting that his edits were recommendations, comments. A lot of them were posed as questions, in fact (p. 152). He also explained that many climate change reports were reviewed and edited by dozens of agencies; Philip saw his edits as just a typical part of the multi-agency process. Regis said, "The issue was all about the editorial process of taking different entities comments and trying to reconcile them. Furthermore, Regis argued that it’s absolutely true that the guy who was coordinating the comments, changed language in the document, but his changes all went in the direction of the actual language given to us by scientists, and then the corrections were reviewed by scientists. In other words, Regis thought Lionel, as a policy person rather than a scientist, should have conceded to the staffer who made the editorial changes (though Regis statement about deferring to the scientists cannot be verified from this research).

Role of politics. A final source of disagreement between the interviewees was the role that politics played in the editorial changes. Lionel noted that his office had the science world on one side, but the White House political world on the other side, and they clashed in that office, which he called a problem in the scientific integrity department. Joan, the non-governmental observer, was more forceful, arguing that the editorial changes happened for political reasons to support a particular position. However, Regis pointed to an investigation by the House Committee on Oversight and Government Reform in 2007, which held hearings on Allegations of political interference with government climate change science. As Regis notes, the committee issued a preliminary report, but not a final report: You’ll find out that there was never any final report issued, because at the end of the day, there was no there, there. From his perspective, political considerations simply did not affect the editing of NACC. Kevin, similarly, did not see any politics involved; he was committed to doing it right, following scientific principles I was interested in doing this, but not if this was going to be a whitewash. Thus, to Kevin, charges of political interference in the NACC or other climate change reports were particularly frustrating.

Analysis of this case. It is unsurprising that Lionel, the alleged victim of censorship, perceived political interference with the NACC. Similarly, it is unsurprising that Kevin and Regis, the alleged censors, perceived the editing to be beneficial and unmotivated by politics. The other differences — the perceived levels of scientific uncertainty and the type of document were unanticipated. Clearly, for Lionel and Joan, the NACC editing was problematic because they viewed it as a science document conveying clear, definitive scientific facts to a broader audience. Just as clearly, Kevin, Philip, and Regis believed these revisions were reasonable changes to a policy document, one which ought to portray the uncertain nature of much of climate science. These core differences cannot be explained simply by referring to the respondents background or training. Lionel had little scientific background; he described himself as a policy person who came at global climate change as a policy problem, while Kevin had a Ph.D. in science, was a professor of environmental studies, and worked for many years on federal environmental issues. In other circumstances, all three of the principal respondents might well have agreed on scientific principles, climate change facts, and potential policy actions. Yet when working on certain climate change documents, particularly the NACC, sharp differences arose.

Case #2: Climate Change Interviews Restricted

This case occurred in an agency in the Department of Commerce (DOC); several agencies within DOC are responsible for studying climate change and global warming. At this particular agency, pseudonymously called Climate Agency (CA), federally employed scientists pursued many different lines of research, but reported that they were blocked from speaking with the media. In some cases, the scientists were told to turn down media requests for interviews or had interviews cancelled on their behalf; in other cases, media contacts were explicitly directed toward scientists who were skeptical of the connections between climate change and human activity.

Four people consented to be interviewed for this case. Nathan was a prominent researcher on hurricanes and the climate; he had to negotiate prior approval before speaking to the media, and several times that approval was denied. David, a public affairs official for CA, was an alleged censor and was effectively responsible for communication between CA scientists and political appointees; approval or
denial of media requests passed through his hands. Natalie was another public affairs official who played a more minor role. Finally, Joshua was a prominent climate scientist, not formally affiliated with the CA or DOC. Through interviews with these four respondents, three foci emerged: the nature of the restriction, the motivation for it, and the outcome. Overall, there was considerable agreement among the interviewees, though some saw more nuances in the situation.

The nature of the restriction. Scientists working for CA were accustomed to speaking freely and frequently to members of the media about their research. Natalie, the public affairs official, explained that reporters for print, radio, and television media were often in direct contact with scientists: a lot of times they’ll just contact our scientists directly, especially if they’ve worked with them before on some other project. As a scientist at the intersection of two complex research areas, Nathan’s work was somewhat esoteric, though his opinion was still sought by the media, particularly after Hurricane Katrina in 2005. Nathan felt that his access to reporters was restricted for political reasons. He said, there was some interference with, from the government, with my ability to communicate with the public. More explicitly, Nathan reported, there was a period when we were having difficulty, where I wasn’t able to do an interview because we had to have interviews approved through CA’s communication office or something, and they were not approved. He recalled specific incidents when he asked for approval to speak with specific media contacts, and the approval was either denied or delayed until the reporter no longer wanted the interview. Joshua, the non-government scientist, noted that he knew Nathan and other scientists had certainly been prevented from giving an interview. Nathan described this as a deliberate maneuver in which the information flow was restricted.

To a certain extent, David, the public affairs official, agreed with the broad contours of what happened. David explained, I would have to go to [my superiors], and say, hey, look, can we allow this scientist to talk to this reporter? He would then relay the response to the scientists and reporters in question. However, David, who was not a political appointee, also saw more nuances in the situation. He called himself a pragmatic advocate for the release of information, meaning that he supported scientists interacting with the media, but realized that more senior political appointees would not approve of open-ended, broad licenses to speak with the media. He explained that he practiced a lot of careful negotiation so that he was not replaced with someone who was more dogmatic, under whom nothing would have come out. Instead, David’s approach was to cultivate trust with the politicals, [then] I was able to get them, to convince them, to open up some of the channels, and let media talk to some of these scientists. While David may have personal motivations for describing the situation as requiring compromise, his perspective also illuminates the complexity of these alleged incidents of political interference. Although portrayed as simple black-and-white travesties by advocacy groups such as the Union of Concerned Scientists, the actual participants reported experiencing more nuance and uncertainty.

Motivations for the restriction. Nathan first suggested that the motivation behind the restricted access was his research: my impression was that they saw some other scientists as having a view which they wanted to have expressed more visibly than mine. I think maybe because they trusted the other scientists’ view. Because Nathan’s research suggested links between climate change and hurricane frequency and intensity, some people thought it supported more aggressive climate change policies (though Nathan himself has never advocated such policies); thus, Nathan thought that his media access was restricted in favor of other researchers whose research was not politically challenging. David viewed the situation similarly, noting that after Hurricane Katrina, the administration wanted one message coming out, and it was all about the recovery effort in the gulf. They didn’t want a lot of distractions, particularly information about how maybe global warming or climate change was making hurricanes stronger.

Both David and Nathan noted that there was some evidence, from media investigations, that approval or denial of interviews was handled by senior political appointees in the DOC. Beyond this broad perspective, though, David did not have much to contribute about the underlying motivations. He explained, I don’t have a lot of insight into motivations. I don’t know exactly what their true beliefs were. Joshua was less circumspect; from his view, it was a political action at the time, and it related very much to the regime of George W. Bush and their suppression of science.

Outcome. All respondents agreed that this situation was resolved satisfactorily: that is, the requirement to seek prior approval for media interviews ceased. Once Nathan spoke publicly about these
restrictions, several media outlets pursued further investigations and Congress held hearings. Nathan explained that this sort of interference through the public relations side of things went away more or less. I didn’t feel any real restriction after that became public. David concurred: Once all of this came out, then it was open air, there was just no more scrutiny, there was no more of this, should I check with this person or with that person first Later, communication policies at CA were further clarified; it was formally stated that scientists did not need to seek prior approval for interviews, but could handle their own media contacts if they wished.

**Analysis of this case.** Some of the respondents have been portrayed as adversarial. However, in these interviews, they frequently voiced similar perspectives and explanations, with more or less nuance. Their respective roles certainly affected how they perceived the situation for example, David saw himself as caught in the middle, advocating on the scientists behalf, whereas neither of the scientists interviewed expressed such a view. On the other hand, all of the respondents described the general situation in the same way, and all seemed pleased with the eventual outcome.

**Findings & Discussion**

Ostensibly, this sort of scientific research conducted by scientists employed by the federal government is created for policymakers. This relationship was taken as a given by most respondents. They assumed that the science being conducted was, or should be, used by policymakers; that was the point, for example, of the documents researched and created by Kevin and his staff. However, the importance of the public as an audience for this scientific research was striking. Seven of the eight respondents explicitly mentioned the public or citizens during their interviews (without prompting from the interviewer). Citizens’ information access was an important concern for nearly all respondents. Speaking of the complicated climate change reports, Nathan emphasized that some assessment reports are designed to provide information to the broad spectrum of people, including policymakers. Notice that Nathan placed the public first, and included policymakers as a subset of the general public. Many scientists viewed their research as being fundamentally for citizens.

Yet, as the previous case descriptions illustrate, citizens access to this scientific research was hindered in various ways. By synthesizing the findings between the two cases, we can see how citizens physical, intellectual, and social access were affected, which in turn has implications for democratic accountability.

**Intellectual Access**

This element of information access is fairly complicated in these cases. On one hand, several respondents indicated that the climate change research in question here contained information that was accessible to a broad swath of people. One of the most important reports, from Lionel’s perspective, was not written in complex scientific jargon and was distributed widely. It went to reporters, it went to educators, it went to scientists, [and] it was posted on the web, explaining the federal capability for climate change research. Nathan told people, if you want to educate yourself, start by reading the summary for policymakers. It’s only, I forget, 20 pages long or something. Every citizen, every policymaker who really wants to truly be informed about this problem should start there. Here, Nathan notes both the availability of the information and the option to read more deeply to access the complex science that underlies the summary report. By directing others to the report, Nathan encouraged their scrutiny and their comprehension, anticipating that this would reduce general uncertainty about climate change science.

However, many respondents were less sanguine about the public’s ability to intellectually access climate change science. One respondent worried about the science illiterate population of contemporary America; others were more circumspect, but nonetheless concerned about how well citizens understood the complex science that shapes much public policy. Joan said, part of the problem is that science is very difficult to communicate scientists are not necessarily good at communication. David concurred with the difficulty of communicating science to the general public, noting that the unfortunate thing is that a lot of times, the situation is so complex, that it’s very difficult to appropriately characterize it and to put it into terms that Joe Public can understand. On the other hand, simplifying too much glosses over the
nuance and complexity of this scientific research. He said he faced a catch-22: you either inundate people with all of the information, or you dumb it down to a point that people can easily ingest it. Many citizens are fairly unknowledgeable about basic scientific information (Stine, 2008; Doremus, 1997), so understanding the complex science of climate change may be particularly unlikely. Joshua noted: there are a lot of polls that have been done and there are things that are demonstrably wrong that are believed by many people in the public. Thus, respondents had mixed views about citizens intellectual access to scientific information. To a certain extent, intellectual access was affected by the restriction of scientific research as described in these cases— for example, the editing changes or the denial of media interviews may have prevented citizens from accessing more easily understood information.

Physical access

In some ways, the restriction of this component of information access is the most clear-cut: when research was edited or scientist denied media contact, citizens did not have physical access to the information in question. The significance of physical access can be seen in the penciled-in editing remarks made by Philip and others; they were literally removing words from citizens access. Lionel called these editorial changes censoring or misrepresenting the communication of climate science.

Nathan, again, neatly summarized the connections between science and the public: In my view, the channels of communication between scientists and the public, through the media, should be open. These channels of communication are physical aspects of information access; these channels are the very conduits that some respondents tried to close.

Social access

In addition to intellectual and physical access, citizens also need social access to scientific research. Social access includes the ways in which one s worldview and normative behaviors affect one s access to information. One way in which social access is implicated in these cases is citizens interest in the scientific research in question. Several respondents indicated that the general public does not seem particularly interested in the intricate details of most science policy.

Joan, for example, said, I think the public, particularly in the last couple of years is much more concerned about other issues than this They re much more worried about the economy and their jobs, all of those kinds of things. The general social climate in the U.S. is not conducive to in-depth consideration and analysis of scientific research. In a more colloquial manner, David attributed the lack of public interest in science policy to the short attention span of the American public. From his perspective, many Americans have too many issues to worry about or are too easily distracted from scientific research and its implications for public policy. In addition, the broad social implications of climate change reports were a significant concern for all respondents. They believed that a strongly worded report, with clear policy imperatives, could have substantial impact on many aspects of modern society.

Implications for democratic accountability

Communicating with the public providing access to the scientific research that guides public policy is seen as important because the public is so important to democracy. Joan thought that better communication of science policy to the public was important because then you would see some feedback into the politicians, I think. She recounted conversations with political leaders who said they just never heard from their constituents on this issue of climate change. In her experience, Congresspeople would say, in disbelief, you re asking me to step out on this issue when there s no one in my state that even cares about it Politicians, elected to represent citizens concerns, were reluctant to act when no concern had been voiced.

Another climate scientist similarly explained that the public influences policy through their voting, so the public is part of the audience for scientific information, because ultimately in a democracy, the voters are in charge. These respondents are directly implicating the theory of democratic accountability: voters have the power to hold their elected officials accountable, but need to have access to information
in order to do so. When their constituents did not express concerns about climate science, politicians did not take action on climate change.

Several respondents hoped that providing more information to the public would result in more interest and passion about climate change public policy, which would then lead politicians to be more active. Citizens need access to the full range of information guiding public policy to gauge the effectiveness and appropriateness of policy actions; armed with such knowledge, citizens can then hold government officials accountable.

Natalie expressed a related aspect of accountability when she discussed the public: The better the general public understands what this agency’s work is doing, which they are ultimately funding, they’ll see the value of their investments in science. Accountability to citizens is important, according to Natalie and others, because taxpayers financially support science policy. This fiscal stake in science policy is both a reason citizens should have access to scientific research, and a reason that science policy should be conducted scrupulously.

**Conclusion**

Among respondents, frustration and concern were strongest when the public’s information access was restricted. Respondents who felt information access was constrained expressed particular concern about citizens’ lack of access. Those who thought information was not restricted, or that the restrictions were insignificant, indicated that citizens could either access the information in other ways or were unlikely to need access to the information. In these cases, the respondents saw a clear connection between citizens’ access to information and broader ramifications of climate change science, such as possibly significant changes to U.S. public policy. For example, Lionel argued that once the impact of climate change became the national conversation, elected officials would be pressured to make substantive policy changes; any action that delayed or prevented the national conversation from occurring such as restricting information access was therefore problematic. Nathan concurred, saying that his role as a scientist is providing it in an understandable form which then leads to the policy question, which is what should humans do about it. The symbolic significance of the public as the legitimate power in a democracy was the basis for both attempts to restrict information access and challenges to those attempts.

These case studies illustrate the utility of the information access framework. By using this framework, we can better understand the ways in which access to information can be facilitated or restricted. The cases also demonstrate the centrality of information access to democratic accountability. Without intellectual, physical, and social access, citizens cannot hold their governmental officials accountable. Dissemination of information enhances democracy.

The theory of democratic accountability is often criticized for being normative and not describing the actual practices of citizens, yet in this research, it does hold some explanatory power. Democratic accountability was the foundation for why scientists wanted their research to be shared with citizens, and was also, perversely, the motivation for restricting access to that research. Further research can examine whether, and to what extent, this is true in other situations. For example, a great deal of research has been done on other complex policy areas, such as the environment and health. Future research could examine citizens’ access to this science, and how access or a lack of access affects democratic accountability. Information access is also important at the state and local levels, not just the federal level. To what extent does access to information about policies (and the underlying information on which those policies are based) facilitate citizen involvement in local and state politics? Additional research is needed to further demonstrate the relevance of the information access framework and the theory of democratic accountability, but this research has established their utility.
References


A Machine Learning-Based Approach to Predicting Success of Questions on Social Question-Answering

Erik Choi  erikchoi@gmail.com  Vanessa Kitzie  vkitzie@gmail.com  Chirag Shah  chirags@rutgers.edu

School of Communication & Information (SC&I)  Rutgers, The State University of New Jersey

Abstract

While social question-answering (SQA) services are becoming increasingly popular, there is often an issue of unsatisfactory or missing information for a question posed by an information seeker. This study creates a model to predict question failure, or a question that does not receive an answer, within the social Q&A site Yahoo! Answers. To do so, observed shared characteristics of failed questions were translated into empirical features, both textual and non-textual in nature, and measured using machine extraction methods. A classifier was then trained using these features and tested on a data set of 400 questions—half of them successful, half not—to determine the accuracy of the classifier in identifying failed questions. The results show the substantial ability of the approach to correctly identify the likelihood of success or failure of a question, resulting in a promising tool to automatically identify ill-formed questions and/or questions that are likely to fail and make suggestions on how to revise them.

Keywords: social Q&A; fact-based questions; machine learning; question success prediction

Introduction

In the recent past, a substantial transformation has occurred regarding people’s information seeking behaviors, especially within online environments. One behavioral pattern that has developed on account of this transformation is the use of web-based question-answering (Q&A) services along with, and often instead of, web search engines. A popular example is Yahoo! Answers, which has over 200 million users and over a billion questions asked, an average of 90,000 new questions per day (Harper, Moy, & Konstan, 2009). These Q&A services typically provide a web-based interface for asking and answering questions in a variety of categories. Questions can be posted and answered by almost anyone, and often there is little to no monitoring or control over users’ activities or quality of content. Such crowd-based Q&A services are often referred to as social Q&A (SQA).¹ Unlike virtual reference (VR) services, which constitute expert based reference interviews conducted by trained librarians via an electronic medium, SQA sites offer very little or no opportunity of interactions between an asker and an answerer to frame the question appropriately. This may result in poor quality of answers or even receiving no answers for a question. For example, Shah et al. (2012) found that within a period of five months, 13,867 questions across the 25 Yahoo! Answers categories were still open to receive a best answer ranking from the original asker, which could be indicative of dissatisfaction with the answers provided.

¹ For a more comprehensive treatment of terminology and typology for online Q&A services, see Choi, Kitzie, & Shah (2012).

Acknowledgements: The work reported here is done under research project Cyber Synergy: Seeking Sustainability through Collaboration between Virtual Reference and Social Q&A Sites, funded by the Institute of Museum and Library Services (IMLS), Rutgers, The State University of New Jersey, and OCLC, Inc. (http://www.oclc.org/research/activities/synergy.html). We also acknowledge Xiao Qin for his assistance with prediction model developments and data analysis. Choi, E., Kitzie, V., & Shah, C. (2013). A machine learning-based approach to predicting success of questions on social question-answering. iConference 2013 Proceedings (pp. 409-421). doi:10.9776/13224

Copyright is held by the authors.
and 4,638 (about 33%) of them did not receive any answers. Since people specify an information need in natural language to others within an SQA site, it is important to investigate how the information need was structured and/or expressed to understand how others interpreted what the original asker intended to look for as compared to the true information-seeking goal. Predicting the likelihood a question failing by determining whether it contains any overarching features of past questions that have failed will help an asker to reconstruct his/her question and increase its potential for success, promoting more effective information seeking behaviors within the SQA context.

The goal of the work is to investigate what makes a question in SQA likely to succeed, defined here as a question that receives at least one answer, or to fail, defined here as a question that does not receive an answer. By looking at questions that fail, examining their shared characteristics and using a quantitative approach to determine the empirical influence these variables might have on question failure, the authors hope to provide a more concrete and robust way to not only identify questions that are likely to fail, but also to provide suggestions and other means for which to increase the propensity for success. In order to accomplish this, an examination of existing works focusing on content-based studies within SQA will be provided in the next section, followed by a method for extracting various features from SQA questions collected from Yahoo! Answers and a technique to build a model that predicts if a question is likely to succeed or not. The model will then be tested for robustness and accuracy, with results being discussed in terms of implications for improvement of SQA services.

**Background**

Within the past few years, various types of social Q&A (SQA) services have been introduced to the public and researchers have begun to evidence interest in information seeking behaviors within these contexts. People ask questions to the community and expect to receive answers from anyone who knows something related to the questions, allowing everyone to benefit from the collective wisdom of many.

These services often supplant search engine use, allowing askers to pose a question in natural language rather than submitting a few keywords to a search engine and to receive personalized answers from other people, as opposed to a list of results. Due to the intrinsic humanistic aspect of the site interactions, SQA outlets pose a benefit to those who may not be finding satisfactory search results using a search engine result page (SERP), and also offer specific social benefits such as the opportunity to solicit and provide opinion and advice-based information, as well as the ability to foster social expression by encouraging users to participate in various support activities, including commenting on questions and answers, rating the quality of answers, and voting on the best answers.

Adamic et al. (2008) found that knowledge resources within SQA comprise a broad range of topics, however are not very deep since many questions asked solicit opinion and advice, while a very small proportion seek fact based knowledge. This observation has been continually made, most recently by Shah et al. (2012), which observed a minor amount (around 5%) of information seeking questions versus advice, opinion or social expression based ones. Further, Agichtein et al. (2008) found that as many SQA sites continue to grow, overall performance in answering fact based questions using traditional relevance measures wanes. This suggests that further studies, such as the one reported here, prove valuable to the field by improving performance on a previously identified weaker facet of the SQA environment and could potentially impact both the types of questions posed in the future, as well as overall community participation and use.

Research on SQA can be divided into two distinct areas of study - user-based and content-based (Shah, Oh, & Oh, 2009). The former examines the factors that comprise interactions within Q&A communities. Shachaf (2010) suggested that while these communities may differ in scope and means of operation, they all operate under the pretense that interaction within an SQA model is multi-dimensional and collaborative, hinging on assessment, motivation, identity formation, and communicative norms unique to this platform. Gazan (2007) performed a content analysis using Yahoo! Answers, dividing askers into seekers and sloths, and concluding that the more active seekers group received a larger proportion of responses than the sloth counterpart. Oh (2012) studied answerer motivations within Health Q&A sites, finding that altruism was the leading factor in answerer participation.

Content-based studies attempt to characterize the components of the actual questions and answers posted to the site. Shah and Pomerantz (2010) identified several textual criteria that comprise a good answer using human evaluators to rank a question on each criteria, while those in the information
retrieval (IR) community use machine extraction methods of textual and non-textual features to predict answer quality (e.g., Text REtrieval Conference (TREC), held annually). One of the overarching conclusions from these studies was that relevance, answer length, presence of outside sources, and time it took to deliver an answer all constitute significant factors in predicting a best answer.

To the best of the authors' knowledge, similar criteria to evaluate the quality of questions asked within an SQA environment have not yet been developed. Instead, most research focusing on questions within this context attempts to classify all questions based on type (e.g. information seeking, advice seeking, opinion seeking, etc.) in order to examine which questions have the best archival value (Harper, et al., 2009). Harper et al. (2009) also distinguish informational questions and conversational questions in order to investigate the level of archival value by exploring the use of machine learning techniques to automatically classify questions. The authors argue that informational questions seeking factual knowledge or objective data in which there exists a “right” answer, are more likely to solicit information that the asker may learn or use, whereas conversational questions, which do not have a “right” or “wrong” answer, stimulate discussion to obtain other people’s opinions or to perform acts of self-expression. Kim, Oh, and Oh (2007) have investigated criteria that questioners may employ in selecting the best answer to their given question. They also studied how types of questions that users ask correlate to these criteria using a data corpus from Yahoo! Answers and found that affective characteristics, such as answerer politeness, tend to matter more for conversational questions, while traditional relevance theory-based characteristics, such as quality and topicality apply more to informational questions (Kim, Oh, and Oh 2007). Their study of 465 queries found opinion seeking questions (39%) to be most frequent, followed by information seeking questions (35%), and suggestion seeking questions (23%). This finding indicates that conversational questions seeking opinions or suggestions are generated more than informational questions within Yahoo! Answers.

Further studies have touched on how examining question types might improve question dissemination among services, predominately within the realm of virtual referencing (VR) (Duff & Johnson, 2001; Pomerantz, 2005; Arnold & Kaske, 2005), however these studies do not directly address specific practical applications for services yielded from the development of such typologies. A typology for classification of failed fact-based questions was reported in Shah et al. (2012) and summarized in Table 1. The authors defined failed questions as those that did not receive a response after three months, a time period by which most SQA community members reported “giving up,” in seeking an answer from the original posted thread. A randomized set of 200 information-seeking questions, defined as questions soliciting a fact-based response, constituted the data corpus.

Findings from the study (Shah et al., 2012) indicate that main characteristics for the 200 failed questions were spread across the categories with significant concentrations in the too complex, overly broad sub-category (68, 34%), followed by lack of information (28, 14%), relatedness (26, 13%), and ambiguity (21, 10.5%) while socially awkward (8, 4%), excessive information (4, 2%), and poor syntax (2, 1%) exhibited a less likely primary influence on failure. Based on these findings, it appears that questions falling within the broader categories of unclear, complex, and multiple questions represent a higher proportion of those that fail in comparison to inappropriate ones, which intuitively suggests that features measuring this latter characteristic may make less of a contribution to the accuracy of the classifier developed within this study.

**Prediction Model Using Automatically Extracted fFeatures**

Although a large number of content based studies within SQA focus on answer quality, as identified by the previous section, there exists a lack of studies examining its counterpart - question quality. Shah et al. (2012) began to address this area by developing a set of characteristics to describe what types of questions fail within an information-seeking context. The current study extends this research avenue by translating these attributes of question failure into empirical features used to develop a prediction model for question failure. In this section, the authors describe a set of experiments that approximate these empirical translations, construct a classifier trained on these features, and test the predictive accuracy of the subsequent model.
Table 1
Typology for failed informational questions developed by Shah et al. (2012)

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Unclear</strong></td>
<td></td>
</tr>
<tr>
<td>Ambiguity</td>
<td>Question is too vague or too broad, and for this reason, is misunderstood or causes multiple interpretations.</td>
</tr>
<tr>
<td>Lack of information</td>
<td>Not enough information exists to identify the asker’s intended information-seeking goal.</td>
</tr>
<tr>
<td>Poor syntax</td>
<td>Question syntax is ill formed, has typos, or has Internet slang that hampers understanding.</td>
</tr>
<tr>
<td><strong>2. Complex</strong></td>
<td></td>
</tr>
<tr>
<td>Too complex and/or overly broad</td>
<td>Question is too complicated and a few people have the ability and/or the resources necessary to provide answers, even though enough details are provided to identify the asker’s intended information-seeking goal.</td>
</tr>
<tr>
<td>Excessive information</td>
<td>Question contains an excessive amount of information that may lose people’s attention to (or interest in) answering it.</td>
</tr>
<tr>
<td><strong>3. Inappropriate</strong></td>
<td></td>
</tr>
<tr>
<td>Socially awkward</td>
<td>Question is inappropriate, too personal, or socially taboo.</td>
</tr>
<tr>
<td>Prank</td>
<td>Question is posed as a joke or to get attention.</td>
</tr>
<tr>
<td>Sloths</td>
<td>Question is homework related and often reflects a perceived “laziness” of the askers to obtain an answer themselves or to actively participate in the SQA community outside of posting questions.</td>
</tr>
<tr>
<td><strong>4. Multiple Questions</strong></td>
<td></td>
</tr>
<tr>
<td>Relatedness</td>
<td>Title and/or content poses more than one question (although they are related), so the answerers may be confused in interpreting the asker’s intended information-seeking goal.</td>
</tr>
<tr>
<td>Un-relatedness</td>
<td>There is more than one question posed and subsequent questions are unrelated, causing potential respondents to be confused in interpreting the asker’s intended information-seeking goal.</td>
</tr>
</tbody>
</table>

Data

A total of 400 questions posed in Yahoo! Answers were used to develop a classifier for this study. This study investigated two sets of questions from Yahoo! Answers - 200 failed, information-seeking questions used in the previous study by Shah et al. (2012), as well as 200 resolved information-seeking questions. Questions defined as resolved were ones in which the asker of a given question selected any answer provided as the best answer that satisfied his/her information need. Both question sets were selected across the 25 Yahoo! Answers categories and collected via the Yahoo! Search Application Programming Interface (API)³.

Extracting Question Features

The current study assumes that the main characteristics of question failure have been identified by the previous study (Shah et al., 2012) and provide several necessary measures that can be translated empirically to construct a model that identifies failed questions. A set of features was selected for extraction in order to address each of the characteristics of question failure developed by the typology, as empirical translations of hypothesized critical variables that influence a question’s likelihood for failure within Yahoo! Answers. Derived from standard data mining approaches, the resulting features identified

³ http://developer.yahoo.com/answers/
best represent the original characteristics developed within the typology, and will now be further discussed.

**Clarity score (ClarityScore).** To quantify the clarity of a question, we decided to employ a query clarity measure often used within the IR domain (Cronen-Townsend, et al., 2002). This measure computes the relative entropy between the query/question language model and the corresponding collection language model. We used the LA Times collection available from TREC with 131,896 documents containing 66,373,380 terms. The clarity score was computed using the Lemur toolkit. This toolkit has been previously used for measuring clarity (see Belkin et al., 2004; Diaz & Jones, 2004; Qiu et al., 2007), including evaluating high accuracy retrieval (Shah & Croft, 2004).

**Syntax (TypoNumber).** Edit distance (Levenshtein, 1966), which compares the common distance between words to the measured distance of the data corpus, as well as spelling, were measured to determine the syntactical appropriateness, and implied resultant clarity, of a question. Misspellings were detected by Jazzy, a Java-based spell checker built on the Aspell algorithm.

**Readability (FleschKincaidReadingEase).** Flesch-Kincaid Readability scores (Kincaid, 1975) were calculated for each question with the hypothesis that a question with an implied higher cognitive load would attract less potential answers, since less community members would be able to understand the information need of the asker. This measure was used to determine complex, ambiguous questions.

**Inverse Document Frequency (iDFCharLength).** Inverse document frequency (IDF) measures were used to determine questions that might be too broad. The authors hypothesized that the more novel terms within the data corpus in relationship to the amount of words contained in a question, the more direct the question was in stating the asker’s information need, and thus, the increased likelihood that the question would be resolved.

**Presence of taboo words (TabooNumber).** Questions were identified as inappropriate by using a dictionary of “taboo” words and assessing whether an identified question within the corpus had any of these defined words. While this measure identifies the theoretical sub-characteristic of taboo and/or socially awkward questions, it does not measure questions that might seek homework help. Therefore future work might look to include a measure that determines whether or not a question directly solicits homework help, perhaps by flagging key words and phrases from questions defined as such. However, this would take time to identify and build a corpus of questions, and as to the best of the authors’ knowledge this corpus is currently nonexistent, so it was not included as a feature for this study.

**Punctuation (QuestionMarkCount).** We identified multiple questions posed as a single information need in a question by counting the presence of a question mark at the end of each sentence within a question posed to Yahoo! Answers, containing a title and/or content. To not misidentify a single question that might have been punctuated with more than one question mark at the end of a sentence in order to emphasize an information need, the technique used only counted one distinct question mark at the end of a word. In order to not confound variables due to the exploratory nature of this study, related versus unrelated content were combined into one categorization.

**Question length (CharLength) (WordCount) (Sentence Count).** Question length constituted a measure of complexity, in which a longer question was hypothesized to correlate positively with question failure since the longer the question, the more cognitive effort needed to process the information need. In addition, a short question might indicate a lack of information provided, which might in turn make it also unclear. The authors measured question length by the number of characters used, the amount of words in the question, and the number of sentences in the content section (if applicable).

**Content (Content).** When posing a question in Yahoo! Answers, there are two fields - question title, where the actual question is posted, and content, where the asker has the opportunity to describe his/her information need further. A question title is required to pose a question, whereas the question content section is optional allows an asker to supply additional information to provide readers with a better understanding of the information need. As the authors hypothesized that presence of content material could be useful in supplying additional contextual information to certain questions, the significance of whether or not a question has content was measured to determine if a relationship existed between presence of such information and whether or not the related question was likely to fail.

---

4 [http://www.lemurproject.org](http://www.lemurproject.org)
Additional Features

Additional textual measures utilized in other works identifying features of questions and/or answers within SQA that affected either question and/or answer performance, were also included to build a more representative model.

**Interrogative words (StartWith).** It is hypothesized that question type might influence likelihood of failure. For example, perhaps informational questions experience more failure than conversational ones. Harper et al. (2009) identified a series of interrogative words (i.e. “who,” “what,” “where,” “when,” “why,” and “how”) that might differ in proportion among informational, or fact-finding questions, versus more conversational ones. The authors found that words such as “where” and “how” were used more frequently in informational questions and “why” in conversational questions. Extending this observation, certain interrogative words might also play a role in contributing to characteristics of failed questions. For this reason, the presence of common interrogative words was used as an additional variable.

**Number of external links (URLCount).** Gazan (2006) divided answerers within the SQA site, Answerbag into two types, specialists who provided answers based on self-identified expertise and therefore did not provide references, and synthesists who provided external sources. He found that members of the Answerbag community rated answers provided by synthesists, containing external sources, higher than those provided by specialists. Based on this observation, the authors decided include number of external links to determine whether or not this factor influenced a question’s likelihood to fail.

The overall results of numeric features (e.g., clarity score, syntax, readability score, inverse document frequency, and presence of taboo words, punctuation, question length, and number of external links) found from failed questions are described in Table 3; the results indicate that inverse document frequency and clarity score are the most significant features of question failure. Additionally, Table 4 illustrates other nominal features (e.g., presence of content for additional information, interrogative words) of failed questions. The results show that more than half of failed questions contain an interrogative word, “what”, followed by “how” and “is”.

Table 3  
Summary of numeric features used from questions.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Min Value</th>
<th>Max Value</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TabooNumber</td>
<td>0</td>
<td>1</td>
<td>0.023</td>
<td>0.148</td>
</tr>
<tr>
<td>TypoNumber</td>
<td>0</td>
<td>49</td>
<td>3.268</td>
<td>3.938</td>
</tr>
<tr>
<td>QuestionMarkCount</td>
<td>0</td>
<td>5</td>
<td>1.408</td>
<td>0.814</td>
</tr>
<tr>
<td>URLCount</td>
<td>0</td>
<td>2</td>
<td>0.035</td>
<td>0.209</td>
</tr>
<tr>
<td>CharLength</td>
<td>19</td>
<td>2840</td>
<td>168.953</td>
<td>205.274</td>
</tr>
<tr>
<td>iDFCharLength</td>
<td>0.204</td>
<td>1.224</td>
<td>0.533</td>
<td>0.134</td>
</tr>
<tr>
<td>ClarityScore</td>
<td>7.284</td>
<td>17.621</td>
<td>11.25</td>
<td>1.307</td>
</tr>
<tr>
<td>WordCount</td>
<td>4</td>
<td>482</td>
<td>31.118</td>
<td>36.936</td>
</tr>
<tr>
<td>SentenceCount</td>
<td>1</td>
<td>81</td>
<td>2.91</td>
<td>4.461</td>
</tr>
<tr>
<td>FleschKincaidReadingEase</td>
<td>-18.2</td>
<td>118.2</td>
<td>76.13</td>
<td>21.087</td>
</tr>
</tbody>
</table>
Table 4
Summary of nominal features used from questions.

<table>
<thead>
<tr>
<th>Feature</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>231</td>
</tr>
<tr>
<td>No</td>
<td>169</td>
</tr>
<tr>
<td>StartWith</td>
<td></td>
</tr>
<tr>
<td>what</td>
<td>240</td>
</tr>
<tr>
<td>where</td>
<td>8</td>
</tr>
<tr>
<td>when</td>
<td>8</td>
</tr>
<tr>
<td>which</td>
<td>0</td>
</tr>
<tr>
<td>who</td>
<td>0</td>
</tr>
<tr>
<td>why</td>
<td>1</td>
</tr>
<tr>
<td>how</td>
<td>35</td>
</tr>
<tr>
<td>is</td>
<td>14</td>
</tr>
<tr>
<td>are</td>
<td>6</td>
</tr>
<tr>
<td>do</td>
<td>8</td>
</tr>
<tr>
<td>does</td>
<td>3</td>
</tr>
<tr>
<td>other</td>
<td>77</td>
</tr>
</tbody>
</table>

Constructing a model with SVM

Methods

Features were extracted from each question using a variety of tools, including Lemur for question clarity; LingPipe, a java natural language processing (NLP) tool (e.g., tokenization, stopwords removal, etc.); and Jazzy, a java-based spell checker. After the features were extracted, a Support Vector Machine classifier was built with split-sample validation and cross-validation using Weka. For further features evaluation, \( \chi^2 \) feature selection method and correlation feature selection method were applied on the entire dataset to weight features and later to reduce the feature vector. More specifically, 66% of the data was used for model training with split-sample validation and a 10-fold cross-validation was performed for robustness evaluation. K-fold cross validation was one way to improve over the split-sample method. The data set was divided into k subsets, and the split-sample method was repeated k times.

Results

Table 5 summarizes the various outcome measures. The measures performed relatively the same, with the highest percentage of accuracy at 77.94% for SVM. Although these percentages are not indicative of a strong model, they represent enough of a difference from the chance levels as can be illustrated by the Kappa statistic, which measures the agreement of predictions with the actual class. The first value of Relative Information Score and Information Score (Kononenko and Bratko, 1991) corresponds to the cumulative information score and the second one corresponds to previous value divided by the number of instances.

Six of the twelve features make contributions to the model with the highest percentage of accuracy at 76.50%; the largest contribution by far made by the feature StartWith, which represents interrogative words as shown in Table 6. The other two features that make significant contributions to the model are IDFCarLength, which represents the number of unique words in the question, and ClarityScore, which represents the complexity of the question (see Table 7). It is also interesting to note

---

6 http://alias-i.com/lingpipe/
7 http://www.cs.waikato.ac.nz/ml/weka/
that questions identified as inappropriate questions, measured here by presence of taboo words, were not found to be prevalent by Shah et al. (2012), yet had a fairly large effect on the performance of the model within this study. This might be due to the methodology employed by the authors (Shah et al., 2012), which coded the corpus using the developed typology on the perceived main characteristic viewed to have the most significant effect on question failure, while potential secondary features were not included.

Table 5
Result of classification on test split.

| Correctly Classified Instances | 106 | 77.94% |
| Incorrectly Classified Instances | 30 | 22.06% |
| Kappa statistic | 0.558 |
| K&B Relative Info Score | 7602.7895 |
| K&B Information Score | 76.0248 bits | 0.53 bits/instance |
| Class complexity | order 0 | 136.0272 bits | 1 bits/instance |
| Class complexity | scheme | 32220 bits | 252.39 bits/instance |
| Complexity improvement | -32083.9728 bits | -251.39 bits/instance |
| Mean absolute error | 0.2206 |
| Root mean squared error | 0.4697 |
| Relative absolute error | 44.1128% |
| Root relative squared error | 93.9206 % |
| Total Number of Instances | 136 |

Table 6
Result of classification on stratified cross-validation.

| Correctly Classified Instances | 306 | 76.50% |
| Incorrectly Classified Instances | 94 | 23.50% |
| Kappa statistic | 0.53 |
| K&B Relative Info Score | 21200 |
| K&B Information Score | 212 bits | 0.53 bits/instance |
| Class complexity | order 0 | 400 bits | 1 bits/instance |
| Class complexity | scheme | 100956 bits | 252.39 bits/instance |
| Complexity improvement | -100556 bits | -251.39 bits/instance |
| Mean absolute error | 0.235 |
| Root mean squared error | 0.4848 |
| Relative absolute error | 47.00% |
| Root relative squared error | 96.95% |
| Total Number of Instances | 400 |
Table 7

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Chi-squared Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>StartWith</td>
<td>122.429</td>
</tr>
<tr>
<td>idFCharLength</td>
<td>68.428</td>
</tr>
<tr>
<td>ClarityScore</td>
<td>20.253</td>
</tr>
<tr>
<td>TabooNumber</td>
<td>9.207</td>
</tr>
<tr>
<td>QuestionMarkCount</td>
<td>8.591</td>
</tr>
<tr>
<td>Content</td>
<td>2.305</td>
</tr>
<tr>
<td>TypoNumber</td>
<td>0</td>
</tr>
<tr>
<td>FleschKincaidReadingEase</td>
<td>0</td>
</tr>
<tr>
<td>WordCount</td>
<td>0</td>
</tr>
<tr>
<td>URLCount</td>
<td>0</td>
</tr>
<tr>
<td>CharLength</td>
<td>0</td>
</tr>
<tr>
<td>SentenceCount</td>
<td>0</td>
</tr>
</tbody>
</table>

Interrogative words might play a role in determining whether a question fails or is resolved by indicating question type, as hypothesized above. In addition, interrogative words could almost represent a clarity measure in the sense that when one of these common interrogative words (e.g. “what”) is used at the beginning of a question, it immediately indicates to the reader something about the nature of what the asker is looking for and how to frame an answer. For example, the word “what” might indicate that the asker is searching for a noun (e.g. What is the capital of France?), whereas the word “how” might indicate that the asker is searching for an opinion and/or directions (e.g. How do you assemble a computer from scratch?). Number of unique words in a question also represents a measure of clarity and also questions identified as too broad, since presence of novel words indicates a question that is more specific and therefore has a clearer identified information need. Finally the clarity score indicates a direct measure of the sub-characteristic clarity. It is interesting to note that these three top features all fit under the “Unclear” major characteristic developed within Shah et al.’s (2012) typology. A secondary feature as indicated by the IDF measure could also be “Too Complex,” although the other feature measuring this characteristic, Reading Level, did not make a significant contribution to the model, suggesting this measure might be representative of clarity. All of the classification results reported in the present section are summarized in Table 8. Classification accuracies of both the model constructed with all question features and the one with selected features are the same. This result shows that six of the twelve features – TypoNumber, FleschKincaidReadingEase, WordCount, URLCount, CharLength, and SentenceCount have no significant attributes to predict the likelihood of failure for fact-based questions in SQA.

Table 8

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Testing</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>400 samples</td>
<td>Split-sample</td>
<td>77.94%</td>
</tr>
<tr>
<td>SVM with selected features</td>
<td>400 samples</td>
<td>Cross-validation</td>
<td>76.50%</td>
</tr>
</tbody>
</table>

Discussion

Limitations

One main limitation of this study lay with the imperfect translation from a theoretical model, exemplified via the characteristics within the typology for failed questions, and the empirical model, or translations of these characteristics into factors that could be measured using text extraction. For example, how can a machine provide a representation of a complex question, given the nuances embodying the concept? The best researchers can hope to do is break apart the characteristics by key facets and identify the appropriate methods and tools by which to define a corresponding feature. For this
reason, future study might focus on using human coders to classify a set of failed questions, using the typology definitions. It should be then be determined whether there exists a significant difference between how questions were classified by humans versus machine classification.

This study is also limited in its generalizability, both because it only samples from one SQA community, albeit a popular one, and since it only samples a small subset of the types of questions that comprise the corpus. However, it can be argued that information-seeking questions have a greater likelihood of being addressed since they actually have an answer. Future study could focus on why questions soliciting more open ended answers may fail, however unless failure is due to question features rather than their actual content, there would be no overarching classifier that could be built to identify the propensity for these types of questions to fail in the first place.

Although the purpose of this study was to determine whether certain observed features could be used to classify failed questions, it did not consider user attributes. Previous work on SQA answer quality has indicated that user attributes make a significant contribution to predicting whether an answer receives a best answer rating or not, perhaps presence of these attributes could also contribute to whether or a question fails (Bian et al., 2008). Since initial findings indicate that the classifier is weak, adding user attributes as a variable in future studies has the potential to improve the performance of the classifier.

Further study should be done in order to improve on the accuracy of the model, as the current accuracy with ten-fold validation is 76.50%. While the classifier developed here included only textual features identified in the previous study (Shah et al., 2012), other research includes such non-textual features of a question and/or an asker to investigate how those features might be related to earning a response. For example, Shah and Pomerantz (2010) included, in order to evaluate and predict answer quality, information from the answerer’s profile, and reciprocal rank of the answer in the list of answers for the given question, and Teevan, Morris, and Panovich (2011) included both properties of the asker including social network use, social network makeup, profile picture, and time of day that the question was posed for investigating factors affecting response quantity, quality, and speed. These works point out that it would be possible for some non-textual features of a question and/or an asker to affect the likelihood of failure for fact-based questions, and therefore a more comprehensive prediction model with both textual and non-textual features may outperform the current prediction model with regard to the likelihood of fact-based questions get resolved in SQA.

Implications

Within an online Q&A platform, machine extraction could be used to measure the degree of existence for variables, as identified in the previous section, which were found to influence variability in a question’s likelihood to fail. Based on the measured identification of these variables, the machine could then employ a pre-identified approach to assist the asker in increasing the likelihood for success. Such approaches could include referring the asker to a different SQA outlet, in which the question has a better chance of getting answered; referring the asker to a VR site where a reference interview can be performed to better elicit and conceptualize an information need; employing an iterative feedback system; and employing an automated measure such as query expansion or syntax correction.

Referring to a different SQA outlet. Different types of SQA outlets exist. Shah, Choi & Kitzie (2012) developed a typology of these outlets into four types: community based, where people exchange information within an online community (e.g. Yahoo! Answers); collaborative, where users can edit the question and/or answer over time to improve it (e.g. WikiAnswers); expert-based, where users receive answers from experts within a specific topic area (e.g. Google Answers); and social Q&A, where people exchange information using their own personal social networks (e.g. Facebook Questions). Presumably each outlet has different strengths and weaknesses in dealing with certain criteria that might contribute to a failed answer. Future study could look at whether the presence of a certain factor predominately factoring in to a question’s likelihood to fail within the context of one type of SQA site could be migrated to another site where this same factor presents less of a likelihood for the question posed to fail.

Referring the asker to a VR site. Expert-based SQA services would provide a viable option here. Aside from addressing questions too complex, VR services could also address questions that may not contain a fully articulated information need since professionally trained librarians could conduct a reference interview with the asker to assist him in fully articulating it (Taylor, 1968). Such questions
probably have not received an answer since it might require someone with expertise in the field to properly understand and address the stated information need.

**Employing an iterative feedback system.** Iterative feedback systems, such as query reformulation in interactive information retrieval (Belkin et al., 2001), have been shown to assist users in articulating their information need depending on what information they initially deem relevant, and how this information is processed by the system to provide better results. Within an SQA context, perhaps archived answers of similar questions could be shown to the user, and the user could pick the answer that is most relevant to his question. If the answer chosen still does not satisfy the query, one of the other suggested methods could then be applied.

**Employing an automated measure.** This would be useful for questions that might lack information necessary to provide a good answer. The system could be trained using simple measures, such as the IDF measure used within this study, to identify questions that might be lacking information and suggest simple techniques such as query expansion or using a thesaurus to suggest more unique terms that might better convey the question to others. Another way to address the problem within the SQA platform is to monitor for syntax and spelling and make suggestions; much like in all commonly used word processing documents, which could go a long way in improving the overall clarity of the question.

**Conclusion**

Predicting the likelihood of failure for fact-based questions in SQA offers a way to assist information seekers in constructing question with an increased likelihood of being answered. The previous study by Shah et al. (2012) attempted to identify several characteristics of why fact-based questions fail in Yahoo! Answers and revealed that the characteristics unclear, complex, inappropriate, and multiple questions are major attributes of fact-based questions that failed within Yahoo! Answers. However, helping an asker revise his/her question might be the first step in making information seeking behaviors more effective in SQA. To do so, it is important to first identify attributes suspected to affect question failure and suggest solutions to improve ways of constructing a question.

Based on a typology for classification of failed fact-based questions (Shah et al., 2012), the study extracted a variety of textual features in order to build a prediction model for determining the likelihood for question resolution. The study found that a question starting with “what”, the number of unique words in the question as measured by Inverse Document Frequency, the level of clarity, presence of taboo words, punctuation, and significance of whether or not a question has content for additional information, are the most significant features for prediction. These findings shed light on the ways in which an asker poses a question and suggests various applications (i.e., incorporating relevance feedback, enabling question routing) that could address how to revise the question in order to increase its likelihood of being answered. Since SQA enables people to seek and share information to fill the knowledge gaps that might not be addressed by other services, providing an appropriate specification and/or structuration of an information need in natural language constitutes a fundamental step toward conceptualizing an effective method for seeking and sharing information. Identifying attributes for why some questions fail and predicting the likelihood of having a question answered presented in the current study will play a significant role in clarifying and revising an asker’s question for a better question-answering process in SQA, and future studies will also benefit from these findings in order to evaluate the quality of questions asked within an SQA environment.

**References**


Retrocomputing as Preservation and Remix

Yuri Takhteyev
University of Toronto
yuri.takhteyev@utoronto.ca

Quinn DuPont
University of Toronto
quinn.dupont@utoronto.ca

Abstract

This paper looks at the world of retrocomputing, a constellation of largely non-professional practices involving old computing technology. Retrocomputing includes many activities that can be seen as constituting “preservation.” At the same time, it is often transformative, producing assemblages that “remix” fragments from the past with newer elements or joining together historic components that were never combined before. While such “remix” may seem to undermine preservation, it allows for fragments of computing history to be reintegrated into a living, ongoing practice, contributing to preservation in a broader sense. The seemingly unorganized nature of retrocomputing assemblages also provides space for alternative “situated knowledges” and histories of computing, which can sometimes be quite sophisticated. Recognizing such alternative epistemologies paves the way for alternative approaches to preservation.

Keywords: retrocomputing, software preservation, remix

Recovering #popsource

In late March of 2012 Jordan Mechner received a shipment from his father, a box full of old floppies. Among them was a 3.5 inch disk labelled: “Prince of Persia / Source Code (Apple) / ©1989 Jordan Mechner (Original).” Mechner’s announcement of this find on his blog the next day took the world of nerds by storm.¹ Prince of Persia, a game that Mechner single-handedly developed in the late 1980s, revolutionized computer games when it came out due to its surprisingly realistic representation of human movement. After being ported to DOS and Apple’s Mac OS in the early 1990s the game sold 2 million copies (Pham, 2001).

Mechner’s original 1989 version, however, was written for Apple II, a platform already somewhat outdated at the time. The original version of the game thus featured much more modest graphics and sound than the later DOS and Mac releases. This early version is still remembered — and played — by aficionados, however, being easily available on the Internet in the form of disk image files derived from a “crack” of the game produced around 1990, credited to “The Crasher” and associates, and bearing a curious dedication to “Nebraska Cement Factory.”²

Prince of Persia images is to load them on one of the many Apple II emulators available online. For the more dedicated fans, however, there is the option of using the original hardware. For some, this original hardware is, of course, Apple II. For others, however, it may be other 1980s computers, including some that could not run the game at the time. For example, in 2011 a programmer known as “mrsid”

¹ Our discussion of Mechner’s discovery and subsequent recovery of Prince of Persia source code is based in part on Mechner’s blog (Mechner, 2012a) and Twitter stream (Mechner, 2012b), as well as the Wired article written by a person who was present during the recovery process (Mastrapa, 2012).

² Our dating is based on the “crack screen,” included in the game which could be a forgery.

Acknowledgements: This research was supported by Insight Development Grant from the Social Sciences and Humanities Research Council of Canada.


Copyright is held by the authors.
successfully completed porting the Apple II version of *Prince of Persia* to Commodore 64, a task that took him two and a half years (mrsid, n.d.). (Mechner’s comments on the announcement noted that the game was not originally released on Commodore 64 because the platform was considered obsolete in 1989.) Projects such as those undertaken by mrsid would be much easier if the source code of the game were available. The code, however, had long been presumed lost. Mechner’s discovery of the floppy thus generated quite a lot of excitement.

*Prince of Persia* images is to load them on one of the many Apple II emulators available online. For the more dedicated fans, however, there is the option of using the original hardware. For some, this original hardware is, of course, Apple II. For others, however, it may be other 1980s computers, including some that could not run the game at the time. For example, in 2011 a programmer known as “mrsid” successfully completed porting the Apple II version of *Prince of Persia* to Commodore 64, a task that took him two and a half years (mrsid, n.d.). (Mechner’s comments on the announcement noted that the game was not originally released on Commodore 64 because the platform was considered obsolete in 1989.) Projects such as those undertaken by mrsid would be much easier if the source code of the game were available. The code, however, had long been presumed lost. Mechner’s discovery of the floppy thus generated quite a lot of excitement.

The find, however, also presented a challenge. “I will now begin working with a digital-archeology-minded friend to attempt to figure out how to transfer 3.5″ Apple ProDOS disks onto a MacBook Air and into some kind of 21st-century-readable format,” Mechner wrote on his blog. Mechner’s call for assistance brought two men to his door a few weeks later. One was Jason Scott, best known as the maintainer of textfiles.com, a website originally dedicated to preserving thousands of ASCII files shared on bulletin-board systems (BBS) in the 1980s and early 1990s, but then expanded to collect shareware CD images, audio files, and other digital artifacts from the era. Until recently Scott pursued those interests while employed as a system administrator. In 2011, however, he joined the Internet Archive to focus his full-time efforts on preservation of old software. The other man was Tony Diaz, a collector of Apple II hardware and maintainer of the website apple2.org, dedicated to photographic images of Apple II. Each man arrived with somewhat different tools. Scott brought DiscFerret, a small open-hardware device designed to read raw pattern of magnetism from a floppy, leaving the analysis and digitization of the pattern to a software tool, thus offering flexible support for a wide range of approaches for storing data, as well as an ability to circumvent many antique copy-protection schemes. Diaz arrived with a van full of Apple II hardware — original, though rigged with substantial hardware and software modifications, including support for Ethernet, not available on the original Apple II.

With their help, Mechner’s original files were transferred to his MacBook Air, in a live-tweeted session tagged “#popsource” that attracted so much attention that Mechner’s website collapsed from the traffic. The source code was then quickly made available on GitHub, a site widely used for sharing open source code. Within hours, GitHub user “st3fan” made a modification commenting out the copy-protection code (st3fan, 2012). This move was purely symbolic, since the posted code was incomplete at the time and could not actually be compiled and run. A few days later, however, a programmer working on an Apple II emulator credited the posted source code as a source of information leading to an improvement in the emulator.³

The story presented above provides a glimpse into the world of retrocomputing, a set of diverse practices involving contemporary engagement with old computer systems. Such practices are primarily private and non-professional, though this is not always the case — there is also a substantial economy providing products and services. And to the extent that retrocomputing participants are “hobbyists,” in the sense of not being paid for their work, they are hardly unskilled amateurs. Rather, their practice often demonstrates deep sophistication. In other words, many of them are “hobbyists” only in the same sense as many of the contributors to open source software, which today underlies much of the world’s computing infrastructure.

Many of the activities that make up retrocomputing can be seen as constituting collection and preservation, and many retrocomputing enthusiasts in fact recognize preservation of computer history as one of their key goals. Such activities involve efforts to collect and restore old hardware, develop

³ The remaining *Prince of Persia* files were posted on Github in May, making it possible, in theory, to compile the code. At the moment it is unclear if anyone has been successful in doing so.
emulators, and build substantial digital collections of old software. For example, it does not take long to find on the Internet disk images for *Prince of Persia* for Apple II, as well as a *variety* of emulators that can run them. Retrocomputing also involves development of sophisticated tools for digitization, such as the DiscFerret tool mentioned above.

At the same time, closer attention to those projects reveals that they cannot be easily understood as just a matter of preservation in the narrow sense of keeping objects from the past fixed in their “original” form. Instead, retrocomputing is often transformative and involves construction of assemblages that “remix” fragments of old systems with newer elements, such as old software running on freshly debugged emulators or original hardware enhanced with contemporary networking. It can also involve a mixture of historic components that were never combined in the past, as in the case of mrsid’s porting of *Prince of Persia* to Commodore 64.

Such “remixes” are sometimes born out of necessity: retrocomputing usually emphasizes active engagement, actually *running* old systems when possible. This sometimes requires replacing pieces that have become unavailable with newer alternatives. Quite often, however, retrocomputing is transformative because simple preservation is not the participant’s only (or even main) goal. Instead, retrocomputing enthusiasts engage with old computing for a variety of reasons, including a simple desire to have fun with old technology. When doing so, they often seek to adapt systems to their contemporary needs — for example, finding ways to connect old computers to newer ones via modern networking.

While such “remixing” may seem contrary to preservation, it fulfills the goals of preservation in a broader sense. In addition to reassembling fragments of computer history into running systems, retrocomputing enthusiasts also do the work of reintegrating fragments into a living contemporary practice. This provides an important set of resources on which more traditionally focused preservation projects can draw. The most obvious of these are the collections of preserved artifacts and tools. Much more important, perhaps, is the retrocomputing ecosystem itself: a persisting system of cultural and economic relationships that keeps alive a circulation of material and digital artifacts as well as deep technical knowledge. It is this system of relationships that ensured that Mechner’s call for assistance was quickly answered by people who arrived equipped not only with the appropriate knowledge but also with the necessary tools.

However, appreciating retrocomputing just for the resources it can provide may miss its deeper value. The practice of retrocomputing also provides space for ongoing circulation of meaning and divergent “situated knowledges,” and for alternative histories of computing. Consequently, it may not only provide us with new insights into the how of digital preservation, but also into the what and why. We therefore need to recognize the value of retrocomputing projects on their own terms and look for ways to provide them with support, while respecting their own objectives.

In the rest of the paper we start by looking at some of the ways in which retrocomputing aligns with the more traditional approaches to preserving the past. We then look at how it deviates from such approaches and engages in “remix.” We argue that such remix may fulfill the goals of preservation in a broader sense. We then turn to the importance of looking at retrocomputing on its own terms, recognizing the divergent situated knowledges embodied in this practice.

**Retrocomputing as Preservation**

Preserving the history of computing can take different forms (Aspray, 1984; Brueummer, 1987). One could focus on preserving the original hardware and original media, perhaps restoring them to the state where they can actually run (Burnet & Supnik, 1996). One can build digital collections of software, to be run either in emulators or on original hardware (Shustek, 2006). One can also collect documents relevant to the history of computing and its historical cultural milieu (Cortada, 2002; Mahoney, 2008). If such documents were born physical, they can be collected in the original physical form or can be digitized. Some such documents, such as text files shared on bulletin board systems in the 1980s, were “born digital” and can be collected as such, perhaps with some adjustments for contemporary encoding methods (Cloolan, 2007). Taking photos of old computers and video recording them in use provides yet

---

4 See also Lowenthal (1989), who argues that preservation should be less concerned with whole, material artifacts, and instead should look for fragments and processes, and Eng-Wilmot (2008) on collage.
another route. All of these approaches are represented in retrocomputing. Without prejudging the relative value of each of these approaches, we focus this section on retrocomputing participants’ efforts directed towards digital preservation of software, in order to document the substantial sophistication of the participants’ preservation efforts.

Emulation of old platforms in software has been widely recognized as one of the promising strategies for digital preservation, and possibly the only viable long term strategy for preservation of software (see Rothenberg, 1995, 1999; Ross 2000). Development of emulators, however, is difficult work requiring substantial expertise both in contemporary software platforms as well as the old systems that are being emulated. For example, developing an Apple II emulator to run inside a web browser requires knowledge of modern JavaScript as well as deep understanding of the workings of Apple II. This work is made particularly challenging by the large number of old platforms. Development of an emulator is also, in an important sense, never quite complete. First, and most obviously, emulators themselves run on platforms that eventually become obsolete, sometimes quite quickly. For example, the ActiveGS emulator for Apple II works with the 2011 version of the Firefox browser, but not with the 2012 version. Additionally, emulating new software sometimes requires adjustments to the emulator due to the imperfect match between the emulator and the original platform. Due to these difficulties, organizing development of emulators as paid professional work can be extremely costly and institutional software preservation initiatives would find it difficult or impossible to commission emulators for all the software that they may want to preserve (Von Suchodoletz et al., 2009).

A distributed, open source style "peer production" approach provides a possible solution. Retrocomputing enthusiasts often have both the requisite knowledge and the interest in doing this difficult work. Their efforts have so far resulted in a large number of emulators, covering a variety of platforms. In the case of relatively popular platforms such as Apple II, one can find emulators that run on Windows, Mac OS X, Linux, as well as the above mentioned browser-based ActiveGS. Less obviously, there are Apple II emulators for platforms that are themselves quite “retro,” such as Amiga OS (Tzvetkov, n.d.) — an example of "remix" to which we turn in the next section. One can also find many emulators for lesser known platforms, such as Apple’s Lisa or the 1974 (pre-Altair) SCELBI-8H.

Some of the emulators are produced by open source projects in the full sense of the term: the code is developed using open source methodology and shared under an open source license. Many are “not quite open source.” For example, a popular MAME emulator is distributed under a license that allows redistribution, but prohibits commercial use, which runs contrary to open source licensing. One of the reasons for this “not quite open source” approach appears to be historical. While retrocomputing in many ways resembles open source and increasingly emulates it, the practice originates from a different subculture, tied more closely to the world of 1980s personal computing, where “shareware” was the dominant form of sharing software. (In contrast, free / open source software’s roots are best traced to academic computing based on Unix.) This slow transition to open source (impeded in part by legacy licensing) highlights the importance of recognizing retrocomputing itself as a historically situated practice.

While use of emulation has often been recognized as one of the main strategies in digital preservation, this approach has multiple problems. It is important to recognize, however, that retrogaming enthusiasts are often quite aware of such problems and seek sophisticated solutions. For example, last year Ars Technica published an essay by “byuu” the anonymous author of BSNES emulator (byuu, 2011). byuu’s essay points to many challenges facing accurate emulation, including, for example, the computational complexity of ensuring the proper timing of emulated instructions. byuu argues that the correct solution would involve re-synchronizing the clocks of the real and emulated processors after each

---

5 Lowood (2011) extends this to screen capture and Machinima.
6 It should be noted that emulators were in use in the retrocomputing community before Rothenberg suggested their use 1995, for example Nutria, a ZX Spectrum emulator developed in 1991 or ZXAM, a Spectrum emulator for Amiga developed in 1993.
7 Development of emulators by retrocomputing enthusiasts is briefly noted by Rothenberg (1999). Other authors often cite retrocomputing emulators to argue for feasibility of emulation (e.g., Ross 2000), but do not usually discuss the provenance of these tools.
8 According to Open Source Consortium’s “Open Source Definition,” an “open source” license cannot restrict the use of software “in a specific field of endeavor,” a clause specifically meant to exclude licenses that prohibit commercial use. Free Software Foundation’s definition of free software also requires that the recipient be allowed to run the software “for any purpose,” and further that he or she should be free to charge others for copies.
9 Retrocomputing is closely linked with the “demo scene” subculture of the 1980s.
instruction, which would require more computational power than today’s hardware can provide.\(^\text{10}\) byuu discusses the tradeoffs between emulation efficiency and accuracy, urging the users to invest in more powerful machines and run more accurate (if slower) emulators, arguing that wide use of power-intensive emulators would help the community to raise expectations for accuracy and make progress towards identifying and fixing discrepancies. byuu’s analysis thus showcases not just sophisticated technical thinking, but also attention to the social elements of the collective projects in which he (or she) is involved.

Emulation requires “images” of original software — files containing the bits extracted from the original media. This would often need to include images of the original system software (for example, Apple II system ROMs), as well as the application software that is to be emulated.\(^\text{11}\) Such images are in fact widely available on hobbyist sites today, often with elaborate metadata. Distribution of such images is, as a general rule, illegal, as it would require permission from the copyright holder, which is usually difficult or impossible to obtain.\(^\text{12}\)

This situation creates a serious challenge for preservation projects undertaken by traditional institutions, which feel that they must keep old software under lock and key to avoid charges of copyright infringement. However, it presents a lesser problem for hobbyists, who are somewhat shielded by their relative anonymity and lack of major assets.\(^\text{13}\) Additionally, retrocomputing participants employ a number of strategies to limit their exposure yet further. One of them is “splitting” the legal risk, by relying on other parties to host the most “radioactive” assets. For example, some Apple II emulators do not include system ROMs but instead provide users with instructions where to obtain them. Such risk-sharing also provides a potential point of collaboration with institutional efforts: institutions may be able to “offload” some legal risk by relying on the retrocomputing community to perform some of these legally risky tasks.\(^\text{14}\)

Disk images collected and circulated by the hobbyists come from sundry sources. Some are actually quite old: the most commonly available image for the Apple II Prince of Persia appears to have originated around 1990. Some are more recent, but still produced by running image ripping software on old machines — or at least older machines. For example, one can read Apple II disks using early PC hardware with special software. This method can be quite challenging due to copy protection, as well as the gradual disappearance of older hardware and knowledge of how to operate it. Perhaps the most sophisticated solution for this problem is exemplified by DiscFerret and KryoFlux. Both are hardware products that sit between a floppy disk drive and a contemporary computer, allowing the latter to scan the raw pattern of magnetization from a disk’s surface, leaving the “parsing” of this pattern to a later step, implemented in software. This allows, among other things, to handle copy-protection methods that rely on specific features of the original hardware to “trick” it (e.g., writing data between the tracks or relying on timing intricacies). Both projects are run by private groups. The more established KryoFlux is proprietary, while DiscFerret is organized as an open source and “open hardware” project.

Running software in emulators is, of course, only one approach. Retrocomputing quite often also involves using the original hardware. While this can mean working with the original media (e.g., the original floppies), this method is problematic since such media can be easily damaged in the process, either through wear and tear or because data may be modified.\(^\text{15}\) The original media is also becoming

\(^{10}\) For an extreme example of this approach byuu cites DICE, an emulator that aims to emulate discrete logic systems (without a CPU) by simulating individual transistor propagation delays.

\(^{11}\) In case of applications, such software serves dual purpose. First, it provides the raison d’être for the emulator. Second, a diverse collection of application software helps improve the accuracy of emulation.

\(^{12}\) This difficulty often has less to do with the copyright holders’ desire to keep old software away from users, and more with the difficulty that aspiring image distributors encounter in even getting a response. In many cases, it is impossible to identify who the copyright holders are, since the companies that produced the software have by now gone through series of mergers and sales or liquidations. In those cases where the current owners can be identified, their legal departments are often reluctant to grant explicit permission, seeing limited upside in doing so, or simply being too busy attending to their current needs.

\(^{13}\) As RIAA’s lawsuits against people participating in MP3 sharing have shown, neither anonymity nor limited assets provide protection against motivated copyright holders. In case of old software, however, the copyright holders are rarely motivated to pursue their rights (and, in fact, may not even be aware that they are the owners of the software).

\(^{14}\) Another potential strategy, though one we have not seen yet realized in retrocomputing, would involve “crowdsourcing” legal research in the same way Wikipedia does with images: countless volunteers take time to identify the copyright status of the images and when necessary engage with the copyright holders to secure proper licensing.

\(^{15}\) Perhaps the most extreme case of this is presented by “Agrippa”: a limited edition diskette released in 1992 containing a poem by William Gibson and designed so that the content of the poem would be irreversibly encrypted after being briefly displayed to the user (Kirschenbaum 2007).
increasingly rare. As a result, old hardware is frequently used in combination with downloaded images. One way to do so involves writing downloaded images to a new floppy disk. This method, however, requires functioning floppy drives, which are among the least reliable peripherals for old computing platforms. A common alternative is to connect old computers to peripherals capable of reading modern media such as SD cards. Such solutions usually involve a combination of hardware and software, which is today available for a number of platforms, including Apple II ("Pseudo Disk II Controller Card," n.d.). The resulting assemblages of hardware and software from different time periods illustrate the issue of “remix” to which we turn in the next section.

Retrocomputing as Remix

While aspects of retrocomputing can be easily recognized as constituting preservation, retrocomputing is often transformative, producing assemblages of physical and digital fragments originating from different time periods and “remixed” in novel ways. Such remix can be done out of simple necessity to substitute missing pieces in order to “restore” a old system to working order. This restoration can be carried out quite conservatively, as is usually the case when it is undertaken by the more traditional memory institutions. For example, when the Computer History Museum in California undertook the project of restoring a PDP 1, effort was made to ensure that any modifications were reversible. Some retrocomputing projects exercise a similar amount of care for preserving the authenticity of restored systems. Quite often, however, they involve creation of assemblages that clearly violate the principles of minimalism in preservation, as illustrated by many of the examples introduced earlier. One can observe cases of old hardware linked to contemporary peripherals, such as SD card readers or Ethernet networking cards. Such hardware may need new software to interact with such devices. Sometimes the original operating system is replaced with something newer altogether. For example, the early 1980s Commodore 64 might be updated to run the 1989 release of JiffyDOS system software, which provides a substantial boost in performance, while some Amiga computers from the 1980s may be setup to run a 2006 version of MorphOS. (Alternatively, an updated version of Amiga OS is also available and can be made to run on 2006 hardware.) Such systems may then be used to run images of software downloaded from the Internet — original, modified (e.g., to remove copy protection), or brand new.

We conceptualize these transformative aspects of retrocomputing as a form of “remix” — a term popularized by Lessig (2008). Like the closely related concept of “collage,” the term “remix” refers to a creative and often playful reassembly of fragments of earlier works into something new. While the reasons for chimeric assemblages described above is sometimes pragmatic, at other times it is simply playful, carried out for fun. At a gathering of Commodore enthusiasts attended by one of the authors, a participant demonstrated an old Commodore 64C system that he had skillfully painted bright blue. He had also hacked it to play stereo sound and wrote his own Commodore 64 software that took advantages of the sound, since none of the original software was designed for stereo sound. When asked about his reasons for repainting the machine, the man explained that the computer’s original white plastic had turned an “ugly” yellow over time. Repainted blue, it looked “awesome.” In other cases, however, the pursuit of fun and beauty often cannot be easily separated from the “pragmatic” motivation for remixing fragments of old computing. Much like Linus Torvalds describing his development of Linux as “just for fun” (Torvalds, 2001), this notion of fun usually implies getting satisfaction in finding solutions to technical problems, thus fusing “pragmatic” and “playful” motivations.

Playful remix inherent in much of retrocomputing may at first blush seem to be in contradiction to efforts preserving the history of computing. This contradiction, however, dissipates with further analysis. Even in the seemingly extreme case of re-painting an old machine to a new color — a step that cannot be undone — the act preserves in that it restores the machine to the “awesomeness” that it once possessed. A machine that was once a source of joy becomes capable of bringing joy once again.

16 Appropriately for our look at retrocomputing, the term “remix” has also come to connote open-ended and distributed reassembly. That is, in a “remix” fragments of the old are often understood to be assembled not to be frozen in a new configuration (as might be the case in a “collage”), but rather to be taken apart and further remixed by others.
More generally, the remix inherent in retrocomputing allows continuous reintegration of elements of past computing systems into an ongoing, living practice. We understand practice as the system of activities comprised of people, ideas, and material objects, tied by shared meanings and joint projects (Takhteyev, 2012; see also Lave and Wenger, 1991). Computing artifacts are born into such systems and have power and meaning because of their linkages to other elements. Over time, however, some elements of these systems disappear or enter into new relationships and abandon the old ones. Commodore International, a company that once made Commodore computers has long since disappeared. Companies that once wrote software for the Commodore have moved on to other projects. Most of the users have also “upgraded” to other computers. Hidden in a basement old computers may remain functional in principle, but become dead in the sense of no longer being incorporated into an ongoing system of activities. With the dissolution of relationships comes the fading of tacit knowledge that once made its use possible (see Galloway, 2011; cf. Collins, 1974; MacKenzie and Spinardi, 1995). Such processes of social decomposition may often be much more damaging to old computing systems than the physical breakdown of the hardware or storage media, and it cannot be stopped by isolating the fragments and shielding them from sunlight or improper temperature and humidity.

The decay of the socio-technical practice, in which the antiquated elements of computing history were once embedded, is partly stopped or even undone in retrocomputing, as ancient fragments are reintegrated into ongoing activities, becoming part of a contemporary living practice. Such integration allows for maintenance (and sometimes recovery) of tacit knowledge. It also makes possible continuous circulation of tools and resources. While retrocomputing is often understood to be the domain of “hobbyists,” it is in fact undergirded by a complex ecology of commercial, hobby, and grey market products and services. Many of the projects mentioned above are undertaken by either for-profit companies or by people and organizations that combine non-profit and commercial activities. It is this complex, interrelated ecosystem that allowed Mechner’s files to be transferred with such seeming ease from the 1980s floppies to GitHub.

One of the surprising (and perhaps disconcerting) aspects of retrocomputing assemblages is the fact that it often involves an admixture of elements from a variety of historical periods: not just from the time of the “original” and contemporary, but everything in between. The transfer of Mechner’s files from floppy disks to GitHub involved not only old and contemporary software and hardware, for example, but also Ethernet hardware and drivers originating from the late 1990s. Someone looking for the Apple II release of Prince of Persia online would likely encounter a cracked version with a dedication to a cement factory in Nebraska and a screen crediting the crackers. This crack — also in some ways a “remix” — appears to be quite old, likely from the early 1990s. Such time-warping, however, reflects the fact that retrocomputing is itself a historically unfolding practice, closely tied to other forms of computing remix. Thus remix may need to be recognized not as corruption of the original but as representing the vibrant nature of computing. When thinking about what it would mean to preserve Prince of Persia, we may need to consider that the version of the game produced by “The Crasher” and his associates may have been played over the years by more people than Mechner’s original, and thus may be as worthy of preservation. Yet more importantly, the cracked version not only preserves the gameplay of Prince of Persia as experienced by many players at the time, it also reflects the legal and cultural battles over intellectual property rights in the 1980s and 1990s, bearing witness to players’ resistance to a property regime that was still fairly new, and painting a richer picture of computing in the 1980s and early 1990s than does the original game.

**Retrocomputing and Situated Histories**

As we have demonstrated above, the efforts undertaken by retrocomputing participants can potentially be of great value for the more traditional efforts to preserve computing history. At the same time, many other aspects of the practice present challenges to conventional preservation, in particular when it comes to what may seem like wanton disregard for authenticity. While authenticity is a thorny issue theoretically, in practice “authentic” archival objects are usually desired: we normally want archived object and records to be properly contextualized and to resemble the original as much as possible.

---

17 This approach to practice also has important parallels to the agency-oriented approach advocated by Dallas (2007).
Retrocomputing, however, presents us with chimeras such as *Prince of Persia* running on Commodore 64, remixing fragments of 1980s computing into something that may *seem* like an authentic artifact from the period to a naive observer, yet in fact presents an entirely novel combination. We may also further worry about the seeming indifference to authenticity exhibited by sites that showcase the “cracked” version of *Prince of Persia* without bothering to even make note of the fact that this version contains modifications added by the crackers. And of course, there is the issue of retrocomputing participants’ focus on the seemingly trivial domains of computing — in particular games — at the expense of the more “serious” ones.18

Our reactions to such chimeras may lead us to educate the participants about proper curatorial practices. Alternatively, we could treat them as a mostly harmless play. We would like to argue, however, that a better reaction may be to use retrocomputing to challenge our own notions of authenticity and to look at retrocomputing armed with Haraway’s notion of “situated knowledges” (1988). Haraway’s approach aims to find a middle ground between a traditional positivist epistemology that seeks objectivity through a singular disembodied “view from nowhere” and, on the other hand, relativism that rejects the notion of objectivity altogether. Instead, in Haraway’s view, knowledge is only available in the form of a multitude of partial and situated perspectives.19

Applying this approach to the history of computing would lead us to avoid measuring retrocomputing against a singular conception of “the history” of computing, but instead to look at the multitude of histories. From this perspective, a Marxist’s history of computing can be contrasted to an engineer’s history of computing, which can be contrasted to that of a gamer (see also Mahoney, 2005). The question then becomes not whether retrocomputing actually preserves the history of computing, but rather whose histories and knowledges are reflected in it, how, and why.

One criticism of retrocomputing that we have mentioned earlier concerns its heavy focus on computer games. While retrocomputing participants do engage with non-gaming applications, games are in fact preserved much more than, for example, business applications and platform. Haraway’s perspective alerts us, however, to the fact that this criticism itself reflects a backgrounding of a situated perspective that pre-assigns value to different domains of computing, deeming some of them as more worthy of serious attention than others. But what if we suspend such assumptions and instead look at the history (and future) of computing through the eyes of people who preserve and play old games? What emerges then is an alternative history of computing, one concerned less with pure engineering innovation and the politico-economic history of the twentieth century, but instead reflecting in much more vivid colors the computer’s changing place in relation to human psyche (see also Turkle, 1984). A proper reading of such history of computing would take seriously the notion of the computer as potentially an “awesome” machine, and may give some credence to the idea that a freshly repainted Commodore playing stereo sound may be more reflective of the 1980s experience of personal computing than a disconnected machine encased in discolored plastic.

Approaching retrocomputing from the perspective of situated knowledges led us to ask what *other* situated histories may be reflected in it. One example we have encountered is the vibrant community of people focused on the history of computer chess. Taking a form of a loose network of websites, the community combines the dedication and zeal that we saw in many retrocomputing subcultures. In fact, our outsider perspective struggled to reduce this community to any simple examples. Nonetheless, “The Spacious Mind” website appears to be characteristic. The site boasts a very large collection (hundreds or perhaps thousands) of chess computing devices and software and corresponding old hardware (“The Spacious Mind,” n.d.). Instead of cataloging the results of human play against chess computers, the chess computer community pits computer against computer in a battle of artificial wits. The Spacious Mind maintains its own records of these games (with interactive playback) but also participates in a global network of computer versus computer chess tournaments, crowning victors in various categories. To keep the matches interesting, the chess computers are typically arranged in to national allegiances for

---

18 While old games are by far the most active domain of retrocomputing, though, many of the other domains are in fact represented, if not as actively.

19 Haraway’s rejection of a single “objective” perspective of reality is distinct from relativism in that it embraces situated perspectives rather than aiming to rise above them. In Haraway’s view relativism and traditional positivism are similar in that “both deny the stakes in location, embodiment, and partial perspective [...] both are ‘god tricks’ promising vision from everywhere and nowhere equally and fully” (584).
appropriate years (usually the year of production), such as the “1980 World HC&CC Championship Revisited Swiss” tournament.

In addition to tournament play, the chess computer community has developed a set of tests to compare chess devices and software. These tests are used to determine if an identical chipset was used by two distinct manufacturers, or if programming was copied from one maker to another. These tests require careful comparisons of timing to test the response to each move, in addition to the actual change of position. Because some chess computers have built-in randomness, other tests force the computer to make certain moves and then, over a two hour game record the moves in an extensive spreadsheet to spot where derivations occur.

To organize this large collection the curator of the Spacious Mind has established a metadata scheme so rich and complex it is hard to imagine that it could arise in any other context. Each record contains 31 metadata fields, two sets of matches against the curator’s collection (often repeated multiple times), two sets of matches against the chess computer community (cross-referenced against other chess computer websites), and a web-based display re-enacting the moves played. Other retrocomputing subcultures also develop their own informal but relatively stable set of metadata fields appropriate to the specific domain — what Bowker and Star would call “local classification schemes” (1999). Such local schemes frequently go well beyond common archive or library metadata.

Recognizing the value of the situated knowledges of the past, as embodied in retrocomputing, does not mean we need to idealize those perspectives. Such situated histories are partial and incomplete. Recognizing the value of partial perspectives, however, points us to a very different set of critiques and solutions: instead of noting the ways in which those histories diverge from “the history” of computing, we may want to ask whose knowledges of the past fail to be registered. (We may note, for example, the shortage of women’s voices.) Additionally, many situated histories may be represented yet easily overlooked among the many others. We can then ask, what can be done to broaden the diversity of perspectives rather than trying to narrow them down.

The approach advocated here has important parallels in some of the new movements in museology, which increasingly recognizes museums as inherently situated agents, reflecting biases and assumptions of specific cultures (Stam, 1993), rather than possessing a privileged perspective of the kind that Haraway would call “a view […] from nowhere.” This perspective is most easily observed in the way institutions increasingly approach artifacts belonging to other cultures, especially former colonies of the West, where the history of eurocentrism is hard to ignore. At a broader level, however, such approaches increasingly recognize the need for more flexible and relational understandings of the notions of objectivity and authenticity. Authenticity is no longer understood to be an inherent or essential part of the object, but rather is “constructed” from patrons’ individual and collective sense of authenticity (Jones, 2010).

Retrocomputing and the Preservation Community

Retrocomputing presents challenges and opportunities to academic and institutional preservation communities. Perhaps the simplest response (other than just ignoring retrocomputing altogether), is to appreciate the technical knowledge possessed by retrocomputing participants and the work they have done, while closing our eyes to those aspects of retrocomputing that may seem to violate the principles of preservation. (After all, in most cases nothing is lost in those remix projects.) This stance would then open the way for traditional institutions to make use of the products of retrocomputing labor and perhaps even steer the participants’ efforts towards projects deemed more important. This approach may in fact be beneficial to institutional preservation efforts and may well be welcomed by the retrocomputing community. It appears that a growing number of institutions are in fact starting to venture in this direction.

20 One response to such prior eurocentrism has involved paying more attention to other cultures “indigenous” curatorial practices around their own objects (e.g., Kreps, 2009). One could note a parallel with our case: retrocomputing participants are often people who have had deep experience with historic computing artifacts back at the time when they were new, either as users or even as their creators. Their ways of handling those artifacts today may therefore deserve attention and respect in the same way as advocated by Kreps. This parallel, of course, is far from precise, since the power relationships in the two cases are quite different.

430
Approaching retrocomputing from the position of what can be gained from it, however, risks missing the deeper value of the practice. We have attempted to demonstrate that remix practices can be understood as a practical consequence of situated knowledges for the various subgroups of participants. Consequently, we suggest that it may be useful for the preservation community to approach retrocomputing from the perspective of Haraway’s epistemology, to embrace new community-driven practices such as remix, and to work with the communities to understand what drives their projects.

Such engagement does not need to be limited to observation. The institutional preservation community (and the iSchool community more broadly) ought to be able to assist retrocomputing projects. Such assistance, however, should start by asking not what retrocomputing practitioners are doing wrong from the perspective of institutional practice or academic scholarship, but rather, looking at practitioners’ own objectives and asking which of those objectives they are failing to achieve and what institutions may be able to offer. For example, instead of asking how private collections of old software could be “ingested” into institutional repositories, it may be more appropriate to ask in what ways retrocomputing practice could be enhanced if we offered the participants access to more powerful infrastructure.

The preservation community (and some of the other subcommunities within the iSchool movement) can help generalize and extend bottom-up practices within the retrocomputing community, function as a liaison between the sometimes distant retrocomputing communities, help bridge their gaps, and adopt common (exchangeable) practices. The challenge with situated knowledge is that it is, precisely, situated, and thus difficult to benefit from broadly. We caution against approaches that attempt to exploit these knowledges, or reduce them to “mere resource” as Haraway warned. We ought to ask, however, what can be done to help such knowledges become more mobile and thus of greater benefit to all.

References


---

21 This approach is now to some extent taken by the Internet Archive, which some other ways straddles the worlds of institutional preservation and retrocomputing.


Chatting with Friends Online While Watching a Video: What and How Much Information Is Retained?

Lin Lin  
University of North Texas  
Lin.Lin@unt.edu

Chris Bigenho  
Greenhill School  
bigenhoc@greenhill.org

Abstract

The increase in connected devices in today’s learning environments places greater control in the hands of students as to how they use their time and attention. The connected devices combined with the popularity of social media and text based chat tools, have made it easier for students to multitask in the lecture-based classes. This paper expands an ongoing study that examines the effects of media multitasking, in particular, chatting while watching a taped lecture, on students’ attention, memory and learning.

Keywords: attention, memory, information behavior, media multitasking, education

Introduction

This study is a continuation of work that examines issues related to media multitasking and learning in educational environments. As university professors and k-12 teachers, we have noticed the rise of connected devices in the classroom and the increased level of multitasking during class lectures. We see students taking notes of the lectures while texting friends on the phone or updating their Facebook pages. In fact, we see students performing these types of dual-tasks throughout the day, which seem to have become a natural part of how they function. However, what is the cognitive cost of this practice?

In an earlier study (citation omitted for blind review), we examined the undergraduate student’s memory recall under nine conditions in a 3x3 study with three levels of environmental distraction and three kinds of note-taking options. We found significant interactions between environment and note-taking method as the computer seemed to mediate the effect of an auditory distraction during the computer aided note-taking. Additionally we found that students were more successful when taking notes at lower levels of distraction and that those who did not take notes consistently performed poorly on the recall tasks.

The result of the earlier study led to this current one, which sought to examine the cost of multitasking in a more authentic environment. Specifically, we wanted to find out to what extent the students’ attention and memory might be affected when they were multitasking between watching a video, taking notes, and chatting with a robot friend.

Theoretical Framework

Poldrack and Foerde (2007) found that people had a harder time learning new things when their brains were distracted by another activity. The Functional Magnetic Resonance Images (fMRIs) used by researchers showed that when people learned without distraction, an area of the brain known as hippocampus was involved. This part of the brain is critical to the processing and storing of information. However, the hippocampus was not engaged when people learned the task while multitasking. Instead, the area of the brain called the striatum was activated. The striatum is activated by stimuli associated with reward or by aversive, novel, unexpected or intense stimuli (Schultz, 2010). Results indicate that learning


Copyright is held by the authors.
while distracted or multitasking alters the brain’s learning processes and changes the way people learn (Poldrack & Foerde, 2007). Foerde, Knowlton, and Poldrack (2006) found that learning new things is dependent on working memory where habit learning is not as sensitive to working memory. Some tasks such as learning new skills may require high cognitive loads, while other tasks familiar and automatic may require lower cognitive loads.

Several recent studies showed that a secondary task might not be detrimental to the primary task (Andrade, 2010; Lin, Lee, & Robertson, 2009). In Andrade’s study (2010), the participants in the experimental group were asked to doodle while monitoring telephone messages for the names of people coming to a party while the participants in the control group were not allowed to do anything but monitor the telephone messages. The doodling group performed better on the monitoring task and recalled 29% more information on a surprise memory test than the group that simply monitored the telephone. The author concluded that doodling as the secondary task may have facilitated the primary task by reducing daydreaming. Lin, Lee, and Robertson (2009) discovered that the students performed better on their reading comprehension when they had TV as background than when they read in silence or read with TV on as a competing task. The TV as a background seemed to have helped the students focus better on their reading. Some new media and literacy scholars have examined the changes or new habits that the intensive immersion of technology may have brought to the younger generations (Gee, 2003; Prensky, 2001; Tapscott, 1998). Prensky (2001), for instance, suggested that the younger generation is comfortable with multitasking, random (vs. step-by-step) and parallel (vs. linear) access to information because they have developed the skills to do so. Stated by Hembrooke and Gay (2003), “the ubiquity, pervasiveness and mobility of new technologies encourage a simultaneity of activities that goes beyond anything our culture has heretofore ever known. Indeed, the ability to engage in multiple tasks concurrently seems to be the very essence or core motivation for the development of such technologies” (p. 1). Neurological work over the last few decades shows that the brain changes based on use (Diamond, 2002). Luciana, Conklin, Hooper, and Yarger (2005) found that the brain’s ability to effectively self-organize competing information remains in the developmental process until 16 or 17 years of age.

This study was conducted in this context. We wanted to examine to what extent our students are capable of switching between different media activities, and when such switching or multitasking activities become detrimental to their attention, memory and learning. The participants were asked to watch a recorded lecture and take notes under two treatments: not chatting or chatting with an Artificial Intelligent (AI) agent named Cheyenne. The following serve as research questions: 1) Is there a difference in the participants’ ability to understand and remember the video content between chatting and not chatting with Cheyenne? 2) To what extent has chatting with Cheyenne affected the participants’ attention, memory and note-taking abilities? 3) What are the patterns of notes and chatting activities when the participants were chatting with Cheyenne, if any?

Methods

Procedure and Participants

Participants were from a mid-sized independent high school in the Southwest United States. An a priori analysis determined that 40 to 54 participants would allow us to detect a medium effect of f = 0.25 with the alpha level set at p < .05. At the time of this writing, we have collected data from 14 males and 13 females with a total of 27 participants. They range in age from 15 to 18 with the average being 17 years and are between 9th and 12th grades. Participant consent and assent forms were obtained for each participant.

We utilized a one-factor repeated measure design with each subject participating in both control (no chat) and experimental (with chat) treatments. To do so, we prepared two different 20-minutes videos of similar length and level of difficulty. The participants were asked to watch the two video lectures in sequence, and told that they would take a multiple-choice quiz on the content after each video. They were asked to take notes on the computer when watching the videos so that they would better understand and remember the lectures afterwards. They were encouraged to take notes in ways that they would normally do, typing words, drawing diagrams or pictures. With one treatment, they were asked to watch the video and take notes only. With the other treatment, they were interrupted or invited by Cheyenne to chat with her. Since we wanted to examine students’ attention and memory of the video lectures in both of the
treatments, one with and one without chatting, we alternated the order of the two videos and chatting options so that the order of videos and chat options would not affect the results of the study. Consequently, we generated four conditions as shown in Table 1 below. The participants were told that their notes, chats, and video lecture quiz results would be collected for research purposes afterwards.

<table>
<thead>
<tr>
<th>Participant Assignments</th>
<th>First Video</th>
<th>Second Video</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition A</td>
<td>Video A (NO Chat)</td>
<td>Video B (Chat)</td>
<td>10</td>
</tr>
<tr>
<td>Condition B</td>
<td>Video A (Chat)</td>
<td>Video B (NO Chat)</td>
<td>10</td>
</tr>
<tr>
<td>Condition C</td>
<td>Video B (NO Chat)</td>
<td>Video A (Chat)</td>
<td>10</td>
</tr>
<tr>
<td>Condition D</td>
<td>Video B (Chat)</td>
<td>Video A (NO Chat)</td>
<td>10</td>
</tr>
</tbody>
</table>

*Note. There are two conditions (Chat/NO Chat) which are presented in an alternate format to help balance any effects of video content and presentation order.

**Instruments and Analyses**

The study employed a mixed method and generated several data artifacts. In addition to the participants’ notes on the video lectures, their chat transcripts with Cheyenne, their post-lecture quiz results, we also asked the participants to respond to an open-ended questionnaire at the end. The questionnaire asked the participant to inform what his or her experience was like in the process so that we could see from the participants’ perspectives to what extent the “chat with Cheyenne” affected or did not affect their ability to take notes and to remember the video. Additionally, the matching video lectures provided a time code that can be aligned with each data type allowing for all data examined across the temporal continuum of the video lectures. Atlas.TI was used to analyze the qualitative data including the participants’ notes, chat transcripts, and survey results. SPSS was used to analyze the video quiz results and examine the possible relationship between watching video with or without chats.

**Results**

The initial quantitative analysis indicated that there were significant differences between the chat (M = 9.59, SD = 2.74), and no chat treatment (M = 11.63, SD = 1.80), t(26) = -3.20, p = .004 (2-tailed) as well as the topic of each video: Video 1 (M = 11.41, SD = 2.50) and Video 2 (M = 9.81, SD = 2.30), t(26) = 2.33, p = .028 (2-tailed). As expected, there were also significant differences in the number of words recorded in notes under the two chat conditions: chat (M = 169.48, SD = 122.63) and no chat (M = 254.00, SD = 132.46), t(26) = -4.97, p = .00001 (2-tailed). The large standard deviation in the finding was likely the result of some students not taking notes at all during the chat condition and the different methods of note taking employed by the participants. Some initial qualitative analyses of the notes and chat transcripts indicated that 1) there were differences in the use of language between notes taking with chatting and without chatting; 2) there were cases of chat transaction ending up in notes; 3) the chat transactions may serve as indicators for the level of engagements. We expect to have the complete data results available and expect to contextualize the quantitative results with qualitative data at the presentation.
Discussion and Scholarly Significance

The initial results showed that performing a secondary task such as chatting during a video lecture 1) may reduce the volume of notes by over 30%; 2) may negatively affect the participants' short-term memory of the video content; 3) may change the notes by adding multiple sources including the chats into the notes on the video content.

The number of portable connective devices continues to increase in classrooms resulting in a need to understand the effects of using these devices while attending to the lectures in the classroom. This includes both educationally designed uses of these technologies for learning as well as the non-instructional student mediated use of these technologies. While the preliminary quantitative results indicate that student's short-term recall was negatively impacted with student mediated use of these technologies, the qualitative data could provide a further window into what is going on behind the numbers and help us understand why some information is retained while others are not as students attend to media multitasking during lectures. We hope that this specific contribution to the field will help provide evidence that can inform the effective use of these technologies as well as lesson design and delivery.

References


Information Recording in Archaeological Practice: A Socio-Technical Perspective

Emad Khazraee
Drexel University
emad@drexel.edu

Abstract

Archaeology is a collective practice and cannot be done in isolation. In addition, given the variety, quantity and scale of archaeological evidence, information technology is a central component of current archaeological practice. This situation provides an excellent case study for the interplay between Information and Communication technologies (ICTs) and institutional and cultural context. This paper reports on a work in progress that examines the role ICTs play in the construction of archaeological knowledge in practice, which focuses on the processes of data recording and information organization. This study uses the conceptual lens of Actor-Network Theory (ANT) and explores important socio-technical aspects of the development of information systems in archaeology. The socio-technical challenges of information recording are conceptualized as practice fault-lines. Three fault-lines of information recording in archaeology are recognized and presented in this paper: 1) Within community vs. cross-community practices; 2) Data management vs. data analysis; 3) Information system designers vs. archaeology practitioners. Recognition of these fault-lines has substantial implications for the design of information organization technologies for collaborative practices.

Keywords: Actor-Network Theory, communities of practice, archaeological recording, inscription, immutable mobiles

Introduction

Archaeology is a collective practice and cannot be done in isolation. The outcomes of archaeological work are the result of collaboration among multiple scholars from different disciplines. Understanding the collaborative nature of archaeology requires an in-depth analysis of the socio-cultural factors of the practice. Moreover, given the variety, quantity and scale of archaeological evidence, information technology is a central component of the current archaeological practice. Therefore, archaeological practice provides an excellent case study for the interplay between ICTs and institutional and cultural context. This paper reports on a work in progress that considers social informatics as “the interdisciplinary study of the design, use and consequences of ICTs that take into account their interaction with institutional and cultural context” (Kling, Rosenbaum, & Sawyer, 2005, p. 6). Also, the study takes a critical orientation to examine the role that ICTs play in the construction of archaeological knowledge in practice.

The relationship between the social and the technical can be conceptualized in many ways (Orlikowski & Scott, 2008). The premise of this study is that the development of information infrastructure is a socio-technical process where work practice and information technology cannot be separated (Monteiro, 2000; Orlikowski, 2007). Given this perspective, the work in progress focuses on the processes of data recording and information organization in archaeological practice.

Conceptual Framework

Using the conceptual lens of Actor-Network Theory (ANT) helps us better understand the entanglement of the social and the technical in work practice. ANT is rooted in the field of science and technology studies (STS). Rather than relying on normative accounts of scientific practice or


Copyright is held by the author.
technological design STS attempts to study how science is actually done and how technological artifacts are actually designed.

ANT considers technoscience processes the result of the alignment of multiple actors (and actor-networks). Therefore, any technological artifact or social order is seen as the result of transformation and alignment of a set of heterogeneous actors into a network (actor-network). ANT provides multiple level of granularity for the analysis of socio-technical phenomena. Thus, every node in a network can be disassembled into another actor-network. Our level of analysis determines where to stop the decomposition.

Four concepts from ANT are of particular relevance to our study: translation (Callon, 1991; Law, 1992), inscription (Akrich, 1992), irreversibility (Callon, 1991; Law, 1992), and immutable mobiles (Latour, 1987). The process by which heterogeneous actors (humans and non-humans) are aligned, and create some sort of order is called translation. Translation "generates ordering effects such as devices, agents, institutions, or organisations. So translation is a verb which implies transformation and the possibility of equivalence, the possibility that one thing (for example an actor) may stand for another (for instance a network)." (Law, 1992, pp. 5–6).

Inscription is the process of translating an intention or interest into a material medium. Inscriptions are "embodied in texts, machines, bodily skills [which] become their support, their more or less faithful executive." (Callon, 1991, p. 143). From the perspective of ANT, design includes translation and inscription processes. According to Akrich (1992), "[a] large part of the work of innovators is that of "inscribing" [their] vision of (or predication about) the world into the technical content of the new object" (p.208). Inscriptions include programs of action which define the roles for users and the systems (Suchman, 2007). "[Inscriptions] are a product of compromise and mutual adjustment negotiated through a series of iterations" (Callon, 1991, p. 143). Programs of action for archaeology are set by the research agendas and the dominant paradigms. One of the goals of the present study is to scrutinize the process of mutual definition and inscription by multiple communities involved in archaeological projects. Examination of how research programs are inscribed into information recording and information organization artifacts is the main thrust of the current study.

Translation and inscription processes go hand in hand. Callon (1991) introduces the concept of irreversibility which explains how inscriptions become durable. Irreversibility refers to the strengths of inscriptions to resist change and explains how one translation can overcome other competing translations. It is important to notice that we are not talking about absolute irreversibility but relatively irreversible translations. Callon (1991) suggests that degree of irreversibility of a translation depends on: 

(a) the extent to which it is subsequently impossible to go back to a point where that translation was only one amongst others; and (b) the extent to which it shapes and determines subsequent translations (Callon, 1991, p. 150).

A final concept from ANT relevant to this study is the "immutable mobile" introduced by Latour. Latour (1987) argues that immutable mobiles are essential knowledge objects of scientific practice. Mobility provides easy transportation of these knowledge artifacts across a network (e.g. a recording system like “Single Context Recording” makes it possible to record the same type of material from different archaeological sites). Immutability keeps the key features of knowledge objects in different places (e.g. same measurements will be recorded for multiple sites and can be used by different scholars in different institutional settings for the same type of analysis). The concept of “immutable mobiles” represents the process of translation and inscription of a research perspective into the knowledge objects that become irreversible. This study considers widely accepted recording methods, recording forms, and categorization systems as immutable mobiles.

Given the ANT lens, archaeological process can be conceptualized as a translation process that inscribes the archaeological context into a narrative; the material context (archaeological site and its recovered contents) is recorded into text (records, drawings, etc), then these texts are used to reconstruct the context in a narrative form (Figure 1). This is a complex process that takes place within a

---

1 A good example of irreversibility is the QWERTY keyboard layout which is not the most efficient keyboard layout considering typing speed. The intent of designer (inscription) was to maintain the typing speed at a certain level to prevent typing bars from jamming. This intent translated into QWERTY layout. Later, the jamming problem resolved but QWERTY layout were become an irreversible translation (Rogers, 2003). Today, there are better competing translations (e.g. Dvorak layout), but it is not easy to use them instead of QWERTY layout.

2 Single Context Recording System is an archaeological recording system developed in 70's and 80's in UK.
heterogeneous socio-technical network which includes human actors\(^3\) (archaeologists and specialists) and nonhuman actors (ordering, recording and measuring devices). This process also requires the alignment of the interest of multiple actants (humans and non-humans).

From the perspective of ANT, the inscription is a cumulative process and each system has multiple layers of inscription. Therefore, we can decompose this heterogeneous network and study the ordering devices within the network, and learn about the inscriptions which are embedded in ordering devices. As mentioned before, research programs or plans of action for research are inscribed into the ordering devices to enable a set of actions and restrict other possibilities. This work in progress follows ANT approach to understand the nuances of data recording and information organization in a multidisciplinary practice such as archaeology.

**Research Description**

This project started in October 2010. Since then, three research sites have been studied by the author: A museum of archaeology and anthropology at a major university in the US, a national archaeological research center in a developing country, and an international archaeological research project. The driving research questions of this research are: 1) *How do archaeologists create their archaeological narratives based on scattered pieces of information they record from the material evidence?* And 2) *What roles do ICTs play in this process?*

This research applies qualitative research methods and ethnographic techniques to achieve a thick description of information practices in archaeology. Data has been collected by various qualitative research techniques such as participant observation, qualitative interviews, think aloud protocols, and document analysis. The analysis has been performed on two levels. The first unit of analysis is individual archaeologists working with other individuals within archaeological groups. The second unit of analysis is archaeological groups (lab and teams) within broader archaeological projects (e.g. a dig site). Qualitative data analysis techniques were then applied to extract the major themes that best present the less-known aspects of information practices in archaeology. Some of the preliminary findings about socio-technical issues of data recording in archaeology are presented here. These findings help us better understand the nuances of information system design for multidisciplinary research environments.

**Discussion**

The process of constructing archaeological narratives is a very complex socio-technical one. Multiple disciplines such as archaeobotany, zooarchaeology, osteology, geomorphology, and art history are involved in narrative construction. The presence of multiple disciplines in the narrative construction process creates boundaries which are also represented in different methods, techniques, research interests, terminologies, and categorization systems which are used by different disciplines involved in

\(^3\) A more precise term in ANT vocabulary is “actant.”
archaeological knowledge construction. In other words, information artifacts are used by established disciplines in archaeology as a means of defining their own territories. However, the outcome of an archaeological project is the result of collaboration among these multiple teams, and the final product of an archaeological project is not simply the sum total of individual reports by each team or specialty. The final narratives are interpretive wholes, outcome of interaction and integration of disciplinary works that explore cross-disciplinary questions. These narratives attempt to depict the practice of a society or settlement from the past in its integrity.

"Getting us (different specialists) together, and having the others as witnesses for what we have said, clearly, has been important to us, and getting people to harmonize the impression that we all take away from the same data. We tried; we did write a paper with five different authors. ... Just one step outside my narrow perception of material they saw something completely different. They were willing totally to overcome the weaknesses of the data and take the general patterns. ... But it was a very provocative moment for me and trying to achieve this goal of integration." [Participant 3]

While production of a cohesive narrative is the goal, there is a tension between different disciplinary territories as they attempt to integrate their outcomes into a single narrative (Figure 2). This tension reveals the existence of multiple fault-lines in archaeological practice. Some of these fault-lines are reflected in the information recording practices in archaeology. Here we present a set of fault-lines observed during our field work.

![Figure 2. Archaeological narrative construction process](image)

**Fault-line One: Within Community vs. Cross-community Practices**

The tension between the established research questions from one discipline and the cross-disciplinary research questions usually creates and forms a fault-line. Each lab in a project has its own established “standardized packages” which enable collective working (Fujimura, 1992). Standardized packages form around the standardized methods developed within a discipline. The accepted research programs of a discipline are inscribed into the recording tools and methods which are used to analyze material culture and remains. In other words, each data recording technology (system) provides opportunities and, at the same time, imposes restrictions which reflect the inscribed research agenda. The description of a cross-boundary research question in comparison to a disciplinary research interest in archaeology is presented below in a discussion by participant 26:

---

4 Evidence from interview transcripts and discussion with the participants are presented in quotation blocks following by participant’s ID.

5 Here, “technology” means devised procedures, forms, categorization schemes not physical devices like scales or calipers.
A cross-boundary research question would be something like investigating the distribution of animal taxa (and/or parts) occurring in various media such as wall-painting, sculptural/architectural features, figurines, and feasting deposits; or co-presence questions such as trying to correlate certain materials with other materials or architectural features in time and space … but within a single lab, they might be more interested in detailed information like manufacture technique or developing a typology. Many labs are interested in sourcing raw material." [Participants 26]

Categorization is a main component of any recording systems. Each categorization system (representing an ontological commitment) provides specific possibilities for analysis. "Standardized packages" of each discipline include generic disciplinary recording technologies which are developed to support specialist, within the borders of a discipline, though not necessarily cross-boundary and multidisciplinary analysis and meta-analysis.

"specialists having developed their databases in terms of their particular research interests and sometimes neglecting to include a more generic description or interpretation of an object in layman’s terms. … when you find objects, sometimes it is difficult to understand what they are when they are described in specialist terminology." [Participants 26]

In the present case, we observed that recording forms, metadata structures, ontologies, and database schemes are the result of the inscription of research programs into information artifacts. As a result, these artifacts which are used to record archaeological evidence, impose a specific perspective on data records and impose certain constraints for interpretation of data (both in present and future).

"when we’re looking at things like, let’s say, a quern stone, people will be obsessed with the geological nature of it, and then the other specialists will be interested in what it was actually used for, so the functionality of it, and then in terms of human remains, you might interested in if you use this quern stone for 12 hours a day for your whole life, what kind of pathologies would you get from it? People have different priorities, and then what normally gets recorded is the specialist’s analysis of it, so you basically find out it’s made of basalt from 15 miles away. That’s it! [Laughter] … but it takes someone from outside of the specialization area to say, “Well, how is that used? And how did they carry that? Because it weighs 25 kilos, and if it comes from 15 miles away, how on earth did they manage that?” [Participants 15]

We have noticed that it is also necessary to record the social context of recording since understanding of social context is crucial for future analysis and the usability of data by people who are outside the project team. For example, it is necessary to record when and why a database scheme is changed. Why a new type of information is recorded. Does it happen as the result of using a new method for analysis? At one site, a new faunal lab director requested a change in the organization of the database to record additional information. Or perhaps the change occurred because of the new cross-boundary research questions that require additional information to be recorded? For example, climate change is a hot research topic in archaeology currently, and it requires collaboration between two distinct discipline-based labs to analyze the floral and faunal remains. Thus, both labs have to support each other's research interests. Our research has not uncovered the use of any efficient mechanism to record the social context of data recording so far. This is an important research area for future study; as such an orientation may contribute to our knowledge of how the social context of recording or “meta-metadata” can be efficiently captured.

Fault-line Two: Data Management vs. Data Analysis

Standard categorization systems can be accepted by wider communities, and these systems can turn into “immutable mobiles” (Latour, 1987). We discussed earlier that Mobility provides easy transportation of these knowledge artifacts across a network of practice while Immutability keeps the key features of knowledge objects in different settings. “Immutable mobiles” help the diffusion of innovation and provide mobility to research programs. This means that a standardized recording system to some
extent enables the recording of similar data sets across a region or a discipline by multiple teams\(^6\). In other words, a standardized recording system can support data interoperability within a community of interest (e.g. zooarchaeology). However, because of the irreversibility of the immutable mobiles and corresponding inscribed research programs, it makes it more difficult for archaeologists and other ancillary specialists to ask novel research questions.

“Making the transfer between these kinds of two or three dimensional arrays (recording forms) to a more of an inventory of information where each bone can have an infinite number of pieces of information, gathered about it, to make it more flexible, took a long time, and I cannot still do things with this data set (recorded in old forms) that I can do with my current data set (recorded in new way) ... in my new strategy, even there I find myself again entering things twice in order to accommodate the new insights that I have. So, after this I decided, no more forms! It's just too confining, it does not let me gather enough information about individuals.” [Participants 3]

Immutable schemes might be desirable from the point of data management. Use of universal schemes helps maintain a reasonable level of data consistency. In contrast, archaeological understanding is an emergent phenomenon that requires dynamic categorizations and evolving recording systems. Sometimes the development of new technologies and research methods requires new types of data to be recorded. Therefore, to support data analysis, recording systems need to be revised very often. This might be at odds with the desire for data consistency in a database.

Fault-line Three: Information System Designers vs. Archaeology Practitioners

Doerr (2009), who actively participated for 15 years in developing an ontology for cultural objects, complains about the proliferation of archaeological categorizations and vocabularies. We also faced the same complaint from computer scientists who noticed that archaeologists are reluctant to accept universal schemes. This causes a tension among the information system designers and archaeologists. The tension is mostly the result of confusion about the aforementioned distinction between data management and data analysis. We should recognize that they are two different processes and none of them can replace the other. Khazraee and Khoo (2011) distinguish categorization for organization from categorization for understanding and suggest that differences between these processes represent the same tension between codification and interpretation.

To provide an example of this situation, we can consider different databases in an archaeological project. The Finds database from one of our study sites works more like a registry for the inventory of archaeological objects and materials from the excavation. It is used to locate the materials in the crates in case that archaeologists need them for further analysis. Also, it provides the connection between each archaeological find and other specialist databases (e.g. Clay object database). Therefore, the role of the Finds database is to support the organization of the material. In contrast, the goal of categorization in specialist databases (human remains, lithics, etc) is to facilitate the analysis of the material. Therefore, we can recognize the two types of databases here as generic databases and analytic databases. Generic data bases are used for data management; and analytic data bases are used for data analysis. Therefore, we should use two sets of design considerations regarding both metadata and database schemes for these different types of databases.

Conclusion

This paper reports on a work in progress that studies archaeological practice. Preliminary findings of this research explore a few socio-technical aspects of the development of information systems in a multidisciplinary practice such as archaeology. Three fault-lines of information recording in archaeology are presented in this paper: 1) Within community vs. cross-community practices; 2) Data management vs. data analysis; 3) Information system designers vs. archaeology practitioners.

\(^6\) Here we are aware that there are other social factors involved in recording process, even in using one standard process by different team, for example Goodwin (1999) reported on the use of Munsell chart by different teams and their differences. But we ignore this case for now to prevent confusion in the discussion.
Different research programs are inscribed into ICT artifacts to enforce certain perspectives into data analysis. While the use of standardized packages within disciplinary labs can satisfy that discipline's research interests, they may fail to answer cross-boundary inquiries. While designing information systems for archaeology, we should be aware of the differences between data management and data analysis. While the former encourages the use of “immutable mobiles” (e.g. shared universal schemes) to support interoperability and long term data stewardship, the latter requires dynamic evolving ontologies to support the process of emerging understanding in practice. Recognition of this fault-line has substantial implications for the design of information organization technologies. A deep understanding of work practice is required to bridge the designer/practitioner gap. Actor-Network Theory has the potential to provide us a framework to achieve such understanding.

References


Abstract

Under a grant funded by the NEH in the US and the AHRC, ESRC, and JISC in the UK, we are collaborating with computer scientists to create digital poetry visualization tools for fellow creative writers and literary scholars, beginning with sonic patterns and moving on first to more abstract figural constructs such as images and metaphors and then to relationships among poems. Like other digital humanities teams, we aim to develop software that will help readers recognize and analyze patterns in and among poems as aids to close readings and eventually to larger scholarly inquiries. What we hope will help distinguish our project from other efforts, though, is the strong emphasis we are placing on poetry's multidimensionality—especially its relationship to and experience of time, which we are working to access through time-dependent visualizations via the metaphor of “flow.”

*Keywords*: poetry, digital humanities, design research, knowledge discovery

Under a Digging Into Data Challenge grant funded by the National Endowment for the Humanities in the United States and the Arts and Humanities Research Council, Economic and Social Research Council, and JISC in the United Kingdom, we are collaborating with computer scientists to create digital poetry visualization tools for fellow creative writers and literary scholars. In particular, we are working to design appropriate visual forms to represent constructs and arrangements in poetry, beginning with sonic patterns and moving on first to more abstract figural constructs such as images and metaphors and then to relationships among poems. Like other digital humanities researchers, we aim to develop software that will help readers recognize and analyze patterns in and among poems as aids to close readings and eventually to larger scholarly inquiries (Chaturvedi, 2011; Chaturvedi, Gannod, Mandell, Armstrong, & Hodgson, 2012; Clement, 2012; Ruecker, Radzikowska, Michura, Fiorentino, & Clement, 2008; Plamondon, 2006; Unsworth & Mueller, 2009). What we hope will help distinguish our project from other efforts, though, is the strong emphasis we are placing on poetry's multidimensionality—especially its relationship to and experience of time, which we are working to access through time-dependent visualizations via the metaphor of “flow.” Throughout this process, we are continually asking whether and how the tool we are developing 1) is genuinely innovative; 2) creates opportunities to produce new knowledge, rather than information only; and 3) can support and enhance the kind of wisdom in literary practice that arises only from close engagement by a reader with the text. The following discussion of our preliminary results shows how we are proceeding in each of these areas.

Our initial focus is on the sonic features of individual poems. Specifically, we are working to reveal the sound structures and moments of phonic changes within a given poem in order to better understand it as a complex dynamic system of sounds. This first stage of our project will equip critics to leverage the close readings they already conduct, opening the poem to new observations and interpretations. At the same time, we take seriously literary critic Marjorie Perloff's warning in her introduction to the 2009 anthology *The Sound of Poetry / The Poetry of Sound* that much recent “empiricist” study allows for little generalization about poetic modes and values: the more thorough the description of a given poem’s rhythmic and metrical units, its repetition of vowels and
consonants, its pitch contours, the less we may be able to discern the larger contours of a given poet’s particular practice, much less a period style or cultural construct. (p. 2)

With this warning in mind, we sought an abstraction or visual metaphor for poetry to help our collaborators create flexible, extensible software that would both enhance and empower close reading and permit useful “generalization about poetic modes and values.” We believe this metaphor will help structure visualizations not only of individual poems, but of larger data sets as well.

In the earliest weeks of our project, we tried and rejected as flawed a number of potential metaphors (a pressure cooker, an expanding universe, a fire, etc). Very quickly, we came to understand that our focus must be experiential; we want the experience we evoke through visualization to be as multidimensional as the experience of poetry itself, equipping users not only to engage writing as text, but also to enhance their engagement with language as experience. Thus, we came to understand that our visualization must itself involve several orders of time and space, and especially that time and temporality in particular are essential to how we conceive of and understand poetry.

Our collaborators’ preliminary visualizations of sound in poetry, as well as examples published by others, did not seem to us to fully embrace this multidimensionality. Early visualizations of Louise Bogan’s poem “Night”—which were really, at that point, as much propositions as our own early metaphors—were flat and static (Figure 1).

![Night by Louise Bogan (Sound Components)](image)

Figure 1. Night by Louise Bogan (sound components). Copyright 2012 by Alfie Abdul-Rahman, Oxford e-Research Centre, University of Oxford, UK. Used with permission.
For example, one visualization helpfully included phonetic characters for each sound and used color-coding to reveal the distribution of vowels and consonants throughout the poem. But it did not show how these sounds connect through and modulate over time. We did not yet see, for instance, visualized relationship between the long “o’s” in “cold remote” and the shorter “oohs” that followed in the next line’s “blue estuaries.” Another early visualization of the same poem was more time-dependent, with arcs showing the frequency and placement of repeated sounds throughout the poem: a method sometimes used to show musical refrains (Figure 2).

Figure 2. Night by Louise Bogan. Copyright 2012 by Alfie Abdul-Rahman, Oxford e-Research Centre, University of Oxford, UK. Used with permission.

This overview image suggests that sonic repetitions are fairly dense in Bogan’s poem, comprising both smaller, fleeting pairings and other, more dominant patterns that link several times throughout the poem. One problem, though, is that it is difficult to see detail. Even if we magnify the image (Figure 3), following individual arcs remains challenging.

Figure 3. Night by Louise Bogan. [Detail.] Copyright 2012 by Alfie Abdul-Rahman, Oxford e-Research Centre, University of Oxford, UK. Used with permission.

Moreover, this visualization raised questions about how rhyme was to be defined and identified, for it did not link “westward,” “estuaries,” “restless,” “inlets,” and other words repeating and reordering the same cluster of sounds. Overall, then, these and other preliminary visualizations prompted us to begin to think more deeply about the experience of the poem and how that experience is created. At this point, we decided that if we were even to begin to articulate and recreate something like the experience of a poem, we needed to include not only sound, but other poetic devices, and we extended our thinking to include the image. Specifically, we have been working to articulate for our computer scientists, for other scholars and poets, and for ourselves what we understand about how sound on the one hand and the image on the other hand develop in time through the language of specific poems. The insights we achieved through this thinking have now led us to a fresh metaphor about poetry and our experience of it: “flow.”

Currently, the team is thinking of the poem as a fluid (or fluids) moving via its linguistic devices and figures through a (self-)defined space. This is not an entirely revolutionary metaphor for poetry; in fact, Alfred Corn points out that both rhythm and rhyme—elemental characteristics of poetry—are etymologically linked to Greek and Latin terms meaning, “to flow” (p. xviii, 72). But this is the first time that the metaphor has been used to develop tools that will allow the scholar actually to view patterns of flow in individual poems. We believe the idea of “flow” is specific enough to be a great help with our immediate
visualization challenge (to see how first sounds and then images move and change through a particular poem). But we are also confident that it will be strong and flexible enough to take on other kinds of problems, both within the poem or in tackling larger, more complex data sets that have not yet been conceived, much less visualized. For example, we are very excited by the possibility of being able to examine how specific patterns of energy (repeating phonemes, for instance) might interact with other patterns (semantic profiles of words, different repeating phonemes or sonic clusters, particular meters, etc.); how they might emerge, intensify, or dissipate through these interactions.

This attention to non-hierarchical dynamism opens new ways both to visualize poetic behavior and understand poetic time. On the sonic level, if various permutations of rhythm and rhyme, including devices like repetition, alliteration, assonance, etc., create distinct measurements and experiences of time, then each poem through its particular flow “shape” creates its own unique temporal experience and identity. In this view, lyric poems might be less manifestations of single, isolated, static “moments” somehow “outside of” time, as they have been seen in the past, than experiences of immersion in multiple temporal flows created, in part, through their sonic and semantic qualities. These durational, shifting patterns may be rendered as various types of flow, such as swells, surface waves, ripples, splashes, eddies, whirlpools, etc.: a flexibility that will enable us not only to differentiate the various flows in a poem, but to see their relative direction, frequency, and power. In response to our thinking in this area, our collaborators are now working to create a model that allows us to view the dynamics of a given poem within a structure that takes the poem’s actual shape.

Unlike other existing programs, our approach, we hope, will not only help users detect, for example, instances of repetition and rhyme and show their spatial proximity within a text. Rather, by introducing temporality, we may also see how those patterns truly change throughout a poem in relationship to other patterns. We may be able to observe modulation from long to short vowels or between full and slant rhymes—or whether and when rhymes reinforce or detract from large temporal thrusts of regular meter. Nor will we be bound to a rigid structure, needing to arbitrarily decide whether to subordinate assonance to metaphor for instance: a problem the developers of the Myopia tool confronted (Chaturvedi et al., 2012). Instead, we will be able to see whether various features converge, whether one flows from another, and which one is dominant in the particular poem in question. This kind of nuance in visual expression and attention to permeable interactions between data and experience fits with the “Humanities Approaches to Graphical Display” recently outlined by Johanna Drucker (2011). We plan to adopt this approach in our explorations not only of sound, but also of image—specifically the “image in time.”

Though the image is often considered to be a medium for introducing stillness into a poem, close analysis reveals that images constructed from language are anything but static. Not only are they apprehended in and across the temporal waves and eddies created by syntax, line, and sentence, they evolve within those temporal movements both as visual and as linguistic objects, accumulating and dissipating presence and immediacy as they simultaneously navigate and work to create the poem’s time. Eventually, we hope not only to use visualizations to observe how these individual images are constructed within and help to create our sense of a poem’s time, but also to perceive how a given poem relates its individual images to each other through repetition and variation across the poem’s temporal as well as its actual space. As with sound, we hope to see, through patterns visually rendered, how these relationships create our experience and apprehension of the poem.

Beyond analysis of individual poems, the metaphor of flow may be productive for scholars researching larger concepts of literary and historic influence and habitual poetic movements that have not yet been conceived—much less visualized—in these terms. Eventually, we hope our software will allow us to advance knowledge in multiple directions by helping to identify larger moments of change and innovation over poets’ careers or even within literary histories. We anticipate being able to recognize, examine, and begin to explain widespread poetic responses to historical events, social phenomena, or technological invention, and to compare works of multiple authors within or across historic periods. In the long run, we may even use this metaphor to help describe psychological and phenomenological engagements with poems and in communities (Csikszentmihalyi, 1991/2008).

So, while our visualizations in this relatively early phase are not yet useful aids to close readings, much less to broad examinations of literary trends, they have pushed us to consider poetic form in new ways and also to understand that to do what we need them to do the visualizations must be not static, as the initial visualizations have been, but time dependent, in motion. Still, the early visualizations and our interactions with them have already changed and deepened our thinking about poetry and the possibilities
presented for the creation of knew knowledge, even new wisdom, in relation to it. We believe that a vital understanding of the human experience of the poem in all its temporal, spatial, historical, and emotional complexity can lead us to a greater understanding of human experience in general.

References


Disparities in Distribution of Information Resources in the City of Philadelphia: Spatial Analysis of Socio-demographic and Information Resource Variables

Emad Khazraee  
Drexel University  
emad@drexel.edu

Adam Townes  
Drexel University  
amt74@drexel.edu

Anthony Grubesic  
Drexel University  
ahq32@drexel.edu

Abstract

This study explored the use of GIS techniques to spatially explore and identify areas where provision of public library and information services are insufficient. By identifying the spatial distribution of inequity of access to information services in Philadelphia. Geographic analysis techniques were applied to investigate the existence of a relationship between spatial orientation, socio-economic factors and information access. The findings of this study suggest the existence of a spatial pattern in relation to disparity of the distribution of information resources in Philadelphia. A secondary analysis utilized the aforementioned findings to assess the existence of patterns of library service provision on a limited data set. The results of analyses indicate that in the disadvantaged areas of Philadelphia, public libraries play the role of information access gateways/bridges. The findings can be used to better locate the future public library and information service centers in the city of Philadelphia.

Keywords: information equity, information access, GIS, geographic information system, urban information gateways

Introduction

The goal of this research is to study the spatial patterns of access to information resources in Philadelphia and the relation of those patterns to the spatial distribution of social and economical factors to suggest the best location for the new information service centers in the city of Philadelphia in response to identified areas of need and inequality.

Information inequality has been an issue of great concern and study for at least 50 years (Yu, 2006). In an historical sense, research into information inequality seemingly coincides with the rise of a postindustrial information society, the civil rights movement and the federal government’s War on Poverty during the 1960’s. Research from late 1960’s through the 1980’s focused on understanding the interaction between information need and some form of disadvantage. Disadvantage was characterized by lack of knowledge in regard to formal information channels, frequent television viewing, sparse usage of reading materials and association with informal information networks that were deficient in the information resources and channels (Childers and Post, 1975; Yu, 2006).

It was during the 1990’s that the concept of a digital divide came to prominence because of the apparent social, economic and racial differences in ownership of home computers and home based Internet connections. More recently topics such as the app gap and the time-wasting gap (Richtel, 2012) indicate that even where disadvantaged populations do have access, pre-existing social, educational and economic problems still influence information seeking behavior and need.

Information behavior is a complex and multifaceted concept. Issues such as inequitable distribution of information resources, lower educational achievement and lack of digital literacy can affect

Acknowledgements: We would like to thank the staff of the Free Library of Philadelphia for their help in providing library statistics and qualitative information in regard to patron usage patterns. Particular thanks to Jeff Bullard, Tamikka Coppin, Chris Kottcamp, Edwin Saeger and Sandra Thompson.


Copyright is held by the authors.
the information behavior of underserved populations. In order to better understand the information needs and policy considerations related to provision of information resources, information access and library services, it is critical to recognize the demographic, social and geographical elements which relate information inequity. This study utilizes a geographic information system to explore the socio-economic and demographic factors of underserved populations in Philadelphia and how they relate to the imbalance in spatial orientation of information resources in the city.

Literature Review

The essential point to be emphasized is that information inequality entails more than simply technologically, socially, economically or institutionally determinant issues, but a combination thereof. Grubesic and Murray (2002) recognized that the digital divide is more dynamic and intricate than many studies suggest. Moreover, it is not simply confined to areas on the basis of socio-economic determinant factors, but is also spatially defined. The spatial orientation of such factors can be tied a users information environment. Agada (1999) demonstrated that an understanding of the information use environment of a user or a group is useful when describing information behaviors of disadvantaged populations. Information use environments can be understood to be the set of elements which affect the availability, access to and use of information (Taylor, 1991).

As it relates to information behavior, lack of information access is an institutional impediment to which disadvantaged populations adapt. Essentially there is an interaction between infrastructural factors such as level of broadband provision or distance to nearest library and information seeking behavior. Sin and Kim (2008) as well as D’Elia (1980) found significant associations between distance from an information resource, such as the public library and use or nonuse of that resource.

The intent of this study was to explore possible correlative relationships between access to information resources and socio-demographic variables that interact at the neighborhood and community level in Philadelphia. GIS analysis provides a powerful tool to graphically and spatially orient the coalescence of infrastructural, environmental and demographic factors. The spatial analysis performed in this study helps to define information use environments in the various neighborhoods and areas of Philadelphia. Spatial analysis is useful for understanding the geographical context of information access issues. For instance one simple and easily mapped factor that can be tied to use is distance to nearest library. Sin and Kim (2008) utilized GIS techniques and found that distance to the nearest public library branch is a statistically significant factor in determining library use by communities. The location of the nearest branch matters in cases where users are not able to afford Internet, or live in areas with low provision.

Koontz et al. (2009) utilized GIS to analyze demographic and socio-economic variables of service areas of closed library facilities in a nationwide survey. The results from Koontz et al. (2009) indicate that the majority of public library closures occur in areas with lower levels of income and education. Consider then, that the presence of a library in a community may represent an enriching factor in a neighborhood information environment. Accordingly, Sin (2010) concluded that the neighborhood information environment can affect information use across individual, household and community levels. By visualizing the existing patterns of information inequity in relation to heterogeneous geographical distributions of wealth, imbalance in distribution of information resources, and socio-demographic factors (e.g. unemployment, household income, and education level) a deeper understanding of communal behavioral norms can be gained. Such an understanding is useful in enacting effective policies and solutions to address information inequity. Understanding the localized problems and association information needs of populations in the underserved areas allows for design and implementation of information resources and services that are relevant and sensitive to communal social norms and needs. Further, development of research on information behavior in relation to the convergence of factors that affect information behavior, within a geographical context has the potential to enhance understanding of general information of underserved populations.
Data Description

Demographic, economic, educational, crime and information access data were collected according to geographic delineations, such as census block groups, census tracts and breakdown by neighborhood. Broadband provision density and sum total of broadband accounts in each block group were also included in the analysis. The data was acquired from various data sources such as the US Census Bureau, Philadelphia NIS CrimeBase, PASDA, Free Library of Philadelphia and NTIA. For this study demographic data (population density and proportion of White or African American population), economic factors (median family income, values of houses, median monthly rent, percentage of vacant properties, and percentage of vacant residential buildings), educational factors (total number of higher education attainment and based on gender) were acquired from the US Census 2000 data. Data for the crime (aggregated assaults, all serious assaults numbers and assault rate per 1000 population) was collected from the Philadelphia NIS CrimeBase. The information about geographic distribution of public libraries was provided by Pennsylvania Spatial Data Access (PASDA). The Free Library of Philadelphia provided the monthly statistics of turnstile, circulation and computer use for nine branches for one year. Finally, broadband provision density and sum total of broadband accounts in each block group of city of Philadelphia in 2009 were acquired from National Telecommunication and Information Association data.

Methods

The goal of this research was to understand the spatial pattern in the distribution of information resources and that distribution’s relation to the spatial orientation of other social and economic factors. Two sets of analyses were performed and are described in the following paragraphs.

Tobler’s First Law of Geography indicates that “everything is related to everything else, but near things are more related than distant things” (Tobler, 1979, 236). In terms of analysis we first conducted an exploratory analysis of the different distributions of demographic, economic, educational, crime and information access data in the Philadelphia area. In this step distribution maps were created to compare the contrasts in different neighborhoods. Then correlation tests were carried out to assess if any of the included socio-economic variables were correlated with each other. Based on exploration and analysis of the geographic distribution of different factors in the city of Philadelphia, the most significant variables were identified for incorporation into the spatial cluster analysis. These variables are Median Family Income, Population Density, Crime Rate, Higher Education and Race (percentage of White or African American population). Spatial cluster analysis is the spatial grouping of geographical units based on their proximity and their characteristics (Jacquez, 2008). "Spatial distributions with values at certain locations showing relationship with values at other locations are named spatial autocorrelation. Spatial cluster is positive spatial autocorrelation when similar values are spatially clustered together" (Lu, 2000). Spatial cluster analysis was utilized to examine whether urban clusters can be defined based on socio-economic factors and the access rate to broadband Internet connections. Spatial cluster analysis was performed with consideration of the spatial proximity of census block groups in relation to major socio-demographic factors mentioned above.

The secondary analysis was more preliminary in nature, and explored the usage rates of nine public libraries in order to assess whether different patterns of use did exist across branches and what the pattern meant when considered in the spatial context of the spatial cluster analysis. Locations of public libraries were also examined in regard to general coverage, to recognize the undeserved areas. Considerations of coverage patterns were based upon the convenient walking distance and speed, which was assessed as a half mile.

---

2 http://cml.upenn.edu/crimebase/
3 http://www.pasda.psu.edu/
4 http://www.ntia.doc.gov/
5 Leon Krier who is a pioneer in urban design defines a convenient walking distance as a 10 minute walk (Krier, 2009). We interpret this notion as a half mile distance. The same metric was used recently by the Philadelphia Research Initiative of the PEW charitable trust to evaluate the coverage of the Free library branches in Philadelphia.
Results and Discussion

The results of this study indicated that Philadelphia is spatially heterogeneous in regard to the distribution of race, income and education level, as well as some disparity in broadband access. The spatial cluster analysis showed that the city of Philadelphia is clustered into five areas. Table 1 represents the socio-economic characteristics and average internet access rates for census blocks in each cluster.

Table 1
Summary of spatial cluster analysis

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Median Family Income</th>
<th>Total Higher Education Degrees</th>
<th>Aggregated Assault Rate</th>
<th>Sum Count of Internet Subscriptions</th>
<th>Percentage of African American Population</th>
<th>Percentage of White Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>117972</td>
<td>444</td>
<td>14.83</td>
<td>49</td>
<td>15%</td>
<td>76%</td>
</tr>
<tr>
<td>2</td>
<td>25308</td>
<td>85</td>
<td>8.52</td>
<td>35</td>
<td>57%</td>
<td>25%</td>
</tr>
<tr>
<td>3</td>
<td>48119</td>
<td>248</td>
<td>3.71</td>
<td>46</td>
<td>9%</td>
<td>82%</td>
</tr>
<tr>
<td>4</td>
<td>30919</td>
<td>138</td>
<td>8.51</td>
<td>34</td>
<td>67%</td>
<td>25%</td>
</tr>
<tr>
<td>5</td>
<td>60363</td>
<td>327</td>
<td>3.75</td>
<td>38</td>
<td>45%</td>
<td>49%</td>
</tr>
</tbody>
</table>

Figure 1 graphically presents the spatial orientation of the five clusters. Cluster 1 corresponds to areas of very high income and high levels of access. Cluster 2 represents a very densely populated, low income low access area. Cluster 3 corresponds to less populated areas of Philadelphia with high income and high access, but not as high as Cluster 1 or Cluster 5. Cluster 4 represents high access, low income, fairly dense population. Cluster 4 probably represents an area with a large population of college students and middle income households. Cluster 5 is composed relatively high income households with high access. Cluster 5 may be constituted of upper middle class households.

When considering Median Family Income, Population Density, Crime Rate, Higher Education and Race (White or African American), Philadelphia has a population distribution that is racially related, at least where spatial orientation, economic status and demographics variables are considered. The results from cluster analysis demonstrated that levels of education and wealth tend to concentrate geographically in particular areas of the Philadelphia. Though not necessarily correlated, these factors did tend to cluster along racial lines and appeared to be based on levels of income and education. Income and education were somewhat positively correlated. The identified concentrations also applied to the intensity of broadband provision and sum count of broadband subscriptions. The areas of actual usage were more closely positioned to the wealthier and more highly educated regions of the city.

The results of the secondary analysis provide findings related to library services. Based on radial half-mile coverage zones of library service it was determined that there is a gap in library coverage in a densely populated and underserved area as indicated in Figure 2 by the shaded circle with the dotted outline. The area around the library (Nicetown Branch) is underserved and highly populated by low income African Americans, with generally low levels of educational achievement. These findings confirm Koontz et al.’s (2009) identification of fewer libraries in poor areas and Sin and Kim’s (2008) findings related to access and use of library services by disadvantaged populations.

Our library usage analysis at this stage was limited to nine branches because at the time of this study just nine branches data was available (library use statistics refer to year 2010; Table 2). Therefore, this analysis is more exploratory in nature to learn whether it can lead to usable findings. Initial results indicate that two of the nine branches (Lehigh Avenue and Haddington branches) act as information gateways. This finding is based on the comparison of turnstile count versus total computer use. In these two branches, the computer use in relation to turnstile is high in comparison with other branches. Information gateway branches are situated on the border between relatively high information access areas and relatively low information access areas, identified in the spatial cluster analysis. Burghardt’s (1971) hypothesis of gateway cities describes a similar phenomenon where people residing in areas adjacent to the divisions between two geographic units, such as neighborhoods and can be associated with access to some resource such as an oasis in an arid area or areas of differing production such as small cities situated on the periphery of rural areas and large metropolitan areas.
The gateway hypothesis is applied here in order to interpret findings regarding libraries situated adjacent to areas of low and high information provision. Residents of the relatively low provision area, utilize the branches in the high access area as a means of bypassing the disparity of access in their own area. Interpretation of the findings based upon library buffer zone and usage analyses verifies the spatial pattern recognized for the digital divide in Philadelphia from the first analysis.

We follow a cross-sectional analysis method for this exploratory study. This study utilizes data set from the Census year 2000 and broadband access in 2009. For future work we will examine spatial patterns of information resource distributions with longitudinal datasets (US Census 2000 and 2010, as well as longitudinal Internet access data) to better understand the dynamics and patterns of information resource distribution in Philadelphia.
Figure 2. Half Mile Radius Library Buffer Zone

Table 2
Summary of the library use statistics of the nine branches of Free Library of Philadelphia

<table>
<thead>
<tr>
<th>Branch</th>
<th>Turnstile</th>
<th>Circulation</th>
<th>Computer Use</th>
<th>Computer use percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastwick</td>
<td>34552</td>
<td>53822</td>
<td>9101</td>
<td>0.26</td>
</tr>
<tr>
<td>Haddington</td>
<td>38996</td>
<td>31507</td>
<td>13689</td>
<td>0.35</td>
</tr>
<tr>
<td>Haverford Avenue</td>
<td>68500</td>
<td>37834</td>
<td>19856</td>
<td>0.29</td>
</tr>
<tr>
<td>Kensington</td>
<td>52669</td>
<td>34514</td>
<td>16523</td>
<td>0.31</td>
</tr>
<tr>
<td>Kingsessing</td>
<td>89374</td>
<td>42948</td>
<td>16069</td>
<td>0.18</td>
</tr>
<tr>
<td>Lehigh Avenue</td>
<td>43997</td>
<td>28434</td>
<td>22889</td>
<td>0.52</td>
</tr>
<tr>
<td>Regional - West Philadelphia</td>
<td>226616</td>
<td>106034</td>
<td>40307</td>
<td>0.18</td>
</tr>
<tr>
<td>Walnut Street</td>
<td>106453</td>
<td>168929</td>
<td>20082</td>
<td>0.19</td>
</tr>
<tr>
<td>Widener</td>
<td>84351</td>
<td>37078</td>
<td>16261</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Limitations

There are two limitations to this study. The first is the limited data from nine of the fifty-four libraries in the Free Library of Philadelphia systems. To verify our findings we need to utilize the complete statistics of all branches. The second limitation, our data about Internet access were limited to broadband accounts and it did not include data about wireless and mobile access. To better describe information poverty we need to use both data about broadband and wireless/mobile access. Recent studies suggest that specific populations tend to access internet through cell phones (Horrigan, 2009). We aim to address these limitations in our future work. Also, it is important to include additional information resources such as community centers, healthcare centers, schools and churches. However, this requires an evaluation of the importance of these centers to access to information.

Conclusion

In brief, the major findings of this study were that the distribution of information resources in Philadelphia are spatially heterogeneous and are correlated with spatial distributions of income level, population density, crime rate, education and race, as well as some disparity in broadband access and library coverage. Spatial cluster analysis shows that city of Philadelphia clustered into five spatial clusters regarding these variables. This study shows that GIS techniques can be used to identify underserved populations in urban areas according to the spatial distribution of socio-demographic and geographic variables. Additionally, this study suggests that two public library branches of The Free Library of Philadelphia, (Haddington and Lehigh Avenue branches) appear to act as information resource gateways. Effectively these branches are located near the spatially defined boundary of high and low information access areas. These branches act as information access entry points for people who live in underserved areas. Branches near these spatially defined boundaries tended to exhibit comparatively higher computer usage statistics in comparison to other branches located in others parts of the city.

Utilization of GIS techniques to understand the provision of public library services can assist in ensuring effective utilization of limited resources. Identifying the correlation of the geographical concentration of socio-demographic factors and the location of information resources has the potential to engender a broader understanding of why historically or traditionally underserved populations have in many cases remained underserved in the contemporary era.

References


Identifying Crossover Documents in an Interdisciplinary Research Environment

Rich Gazan
University of Hawaii
gazan@hawaii.edu

Abstract

AIRFrame is a NASA project to analyze and integrate astrobiology documents from diverse disciplines to catalyze new knowledge. This paper outlines the technical infrastructure of the current system and reports on an ongoing iterative evaluation, to address the question of how scientists perceive and integrate crossover documents in their research. Some of the obstacles preventing AIRFrame from gaining traction with its target audience of astrobiology researchers include representing their research output accurately, effectively translating and relating diverse metadata, and understanding disciplinary norms and the broader knowledge production infrastructure. The skills required to address these needs suggest a role for both researchers and information professionals to work in tandem with technical tools to catalyze interdisciplinary knowledge. A graduate seminar in interdisciplinary knowledge production, targeted at both researchers and graduate students at the University of Hawaii, has been designed to elicit and impart needed information as input to ongoing AIRFrame development.

Keywords: astrobiology, interdisciplinarity, scientific communication

Introduction

Traditional inputs to scientific research have been largely limited to discipline-specific scholarly publications and datasets, but integrative science requires that information be translated and transported across fields, and applied to common problems. This requires not just technology-based support systems, but nuanced understanding of the knowledge production infrastructure in which scientific research takes place. This paper discusses the iterative evaluation of the Astrobiology Integrative Research Framework (AIRFrame), a NASA Astrobiology Institute (NAI) project designed to identify particular documents from an astrobiology corpus that might be used to connect and mutually inform researchers from different disciplines, and how scientists perceive and integrate crossover documents in their research. The development of a seminar designed to instill interdisciplinary practices and perspectives in both researchers and graduate students is proposed in response to the findings reported here.

Background

Interdisciplinary research is notoriously difficult to quantify (Morillo et al., 2001). Heuristic indicators such as co-authorship or joint project participation are often used to assess interdisciplinary science (Rafols & Meyer, 2010), but the extent to which new knowledge is actually catalyzed as a result of these interactions is open to question. Scientometric analysis of research documents has developed in response to the need to quantify and evaluate the actual and potential impact of scholarly works (Huutoniemi et al., 2010).
Cyberinfrastructure research and education has devoted significant attention to the “last mile”— the last section of connectivity from the network to the user (Gabridge, 2009). Equally important is the “first mile”: the range of inputs available to a scientist conducting research. Scientists are traditionally trained to draw from data sources within their discipline to ground their contributions; to do otherwise risks the perception that the contributions are less relevant or valid. However, this insular approach can work against boundary-crossing research. To do integrative work, researchers must have the means to develop a working understanding of the knowledge production infrastructure within which their research takes place, in its technical, logical and social dimensions (Friedlander, 2008).

The NASA Astrobiology Institute (NAI) consists of 18 teams and over 800 researchers studying the possibility of life beyond earth. This is an inherently interdisciplinary endeavor, with astronomers, biologists, chemists, engineers, hydrologists, meteorologists, oceanographers and researchers from many other fields addressing different aspects of the same question. One of the primary goals of NAI is to catalyze interdisciplinary knowledge, and it provides a hospitable environment for exploration beyond one’s home discipline.

Our approach has been to develop a set of tools and methods to identify potential “crossover” documents within astrobiology, the boundary objects (Star & Griesemer, 1989) across which diverse scientists might communicate and inform one another. To catalyze interdisciplinary knowledge in astrobiology, we have developed the Astrobiology Integrative Research Framework (AIRFrame), a conceptual framework and set of tools and methods for collecting, analyzing and integrating astrobiology documents (Gazan, 2010). Our approach has focused on textual analysis of documents from diverse disciplines to reveal implicit relationships.

Our previous work (Gowanlock & Gazan, 2012) represented the research tracks of scientists at the University of Hawaii NASA Astrobiology Institute (UHNAI) over a ten-year period. Documents were represented by their abstracts and the abstracts of the publications they cite, and their source discipline by the Journal Subject Category from the Thomson Reuters Web of Knowledge database suite. Using an unsupervised machine learning clustering technique, the sequential Information Bottleneck (Slonim et al., 2002), each paper was assigned a cluster, where publications in the same cluster share mutual information. The publications of most UHNAI researchers clustered with those in similar disciplines across multiple clustering runs. However, some publications were found in multiple clusters, and some authors’ work consistently clustered with that of non-obvious colleagues. The clustering and classification processes yielded the first step in a data-driven analysis designed to identify actual and potential crossover points for the integration of diverse knowledge.

Method

While much of our initial work has been accomplished via automated document harvesting from databases such as the NASA Astrophysics Data System (ADS), and clustering using the Weka machine learning and data mining toolkit (http://www.cs.waikato.ac.nz/ml/weka/), integrating documents from other databases presented challenges that a purely algorithmic approach could not solve, motivating the present study. Critical tasks such as searching both Web and professional databases for relevant astrobiology literature, and interpreting, creating and equating diverse metadata, required the participation of individuals with integrative information skills, including researchers, graduate students and others supporting their research, warranting a focus-group data collection approach to iterative evaluation including stakeholders from all these groups.

One of the outcomes of AIRFrame will be a system that presents researchers who seek to do interdisciplinary science with a suggested set of related papers outside their areas of expertise— essentially a recommender system for crossover documents. But like any recommender system, user feedback is critical. In the initial phase of iterative evaluation, members of the UHNAI team were presented with a summary of the results published in Gowanlock and Gazan (2012) in a weekly astrobiology seminar, and participated in open-ended questioning and discussion about their reaction to the design and functions of the system, and recommendations for its productive evolution.
Results and Discussion

Researchers’ responses included concerns about how accurately the document abstracts and abstracts of cited articles represented their research areas and points of potential crossover. Some were pleased at the potential interdisciplinary breadth of their work suggested by AIRFrame, while others questioned why the system had not represented tangible points of crossover they knew to be present in their work. Researchers were made aware of the probabilistic and relativistic nature of clustering, and that documents might cluster differently given the content of other documents in the analyzed corpus. The extent to which users understand and trust the system inputs and representations is understandably related to their likelihood of using the system.

Another weakness of relying on published works is that the research tracks represented are necessarily backward-looking, when the goal is to generate useful crossover documents to catalyze interdisciplinary knowledge for present and future research. Indeed, several researchers found that papers they had published in areas entirely unrelated to astrobiology had been harvested and used to represent their interdisciplinary potential in the field.

The ability to curate data from diverse sources, distinguish relevance, and understand the nuances of different disciplinary literatures and publication norms are some of the critical skills necessary to represent documents and authors accurately, and to normalize representations of their work for comparative analysis. The University of Illinois recently introduced a Data Curation Education Program within the Graduate School of Library & Information Science that integrates many of these skills: “Data curation activities enable data discovery and retrieval, maintain data quality, add value, and provide for re-use over time.” (http://cirss.lis.illinois.edu/CollMeta/dcep.html)

To this end, a seminar in interdisciplinary knowledge production is in development at the University of Hawaii, which will engage graduate students and researchers across campus, and cover the following topics:

- The global information infrastructure
- Scientific knowledge production
- Scientometrics
- Disciplines and interdisciplinary collaboration
- Intercultural communication
- Data mining
- Searching across disciplinary literatures
- Ownership of scientific information
- Data access and visualization
- Science policy
- Economics of science
- Citizen science

Where a generic seminar offering might be easily ignored by overcommitted researchers, especially one outside their domain of interest, the institutional infrastructure of the NASA Astrobiology Institute requires that funded participants submit regular reports on their progress, and that they specifically address how their work crosses disciplinary and institutional boundaries. Thus, NAI researchers are highly motivated to learn how they can maximize and articulate the interdisciplinary aspects of their work. Existing astrobiology seminars are one form of boundary object where diverse researchers share information with students and one another, and adding components of interdisciplinary knowledge production provides a more explicit mechanism to promote meaningful and actionable interdisciplinary collaboration.

The seminar addresses researchers’ roles as generators, seekers, consumers, integrators and stewards of scientific information (ARL 2006). Information that is inaccessible, untranslatable, informal or simply lost cannot be used to represent and suggest areas of potentially productive crossover. The source of these hidden representations is the researcher, but eliciting them often requires the contributions of information professionals. Unpacking the affordances and constraints of the research infrastructure necessitates a sociotechnical approach, and will allow the next generation of researchers and information professionals to confront and address issues of data management, intellectual property rights, data
representation and integration, interdisciplinary and international collaboration, and the publication, dissemination and preservation of scientific information.

Modern researchers, and those who help provide access to the products of their research, must incorporate and critically evaluate relevant data from diverse sources. Works published in non-core databases as well as unpublished and informal documents also represent science practice, and should form part of any system designed to measure and promote interdisciplinarity (Sugimoto 2011). Individuals with skills of data curation and integration will be well positioned to develop not only a broad, systematic view of scientific practice, but be in a position to identify, create or act as boundary objects in future interdisciplinary efforts.

**Conclusion**

The crossover documents suggested by AIRFrame are designed to increase the interdisciplinary aspects of current and future research, but might also be used to inform the development of astrobiology syllabi and curricula, to train the next generation of astrobiology researchers. AIRFrame can also be used to measure the breadth of disciplines from which papers and researchers draw, allowing fine- grained analysis of interdisciplinary work. Through iterative evaluation, the results are shared with researchers, providing them a working understanding of some of the scientometric indicators by which their work will be evaluated. Understanding how research is represented and transported in formal and informal documents, and some of the barriers researchers face to interdisciplinary work, will allow more accurate representations of scientists’ work, and the identification of more potential points of crossover to catalyze cross-disciplinary research.

**References**


Are You Satisfied Yet?
Shared Leadership, Trust and Individual Satisfaction in Virtual Teams

Lionel Robert
School of Information
University of Michigan
lprobert@umich.edu

Sangseok You
School of Information
University of Michigan
sangyou@umich.edu

Abstract

Virtual teams provide organizations with numerous advantages by allowing them to assemble individuals irrespective of their physical location. Unfortunately, dispersion and reliance on virtual communications are two characteristics of virtual teams that can reduce team member’s satisfaction. Promoting satisfaction is important because it is associated with increases in individual behaviors that promote team performance. To address this issue, this paper conducts a multi-level analysis that investigates the influence of shared leadership and individual trust on individual satisfaction in virtual teams. Results indicate that both shared leadership and individual trust increase individual satisfaction. In addition, the significance of a cross-level moderation effect between shared leadership and individual trust indicates that the two act as substitutes for one another. In essence, either shared leadership or trust can be used to facilitate individual satisfaction in virtual teams.

Keywords: virtual teams, leadership, trust, information systems, computer supported cooperative work

Introduction

Virtual teams provide organizations with numerous advantages by allowing them to assemble individuals irrespective of their physical location (Robert, Dennis, & Ahuja, 2008). Despite this, virtual teams also present new challenges (e.g., Alnuaimi, Maruping & Robert, 2009; Alnuaimi, Maruping & Robert, 2010; Dennis, Robert, Curtis, Kowalczyk & Hasty, 2012; Jarvenpaa & Leidner, 1999). Promoting and maintaining a positive collaborative experience in these virtual teams are two such challenges (Morris, Marshall & Rainer, 2002). Dispersion and the reliance of virtual communications are associated with negative team interactions which reduces team members’ satisfaction (Chidambaram, 1996; Kayworth & Leidner, 2000). Finding ways to promote satisfaction is important because it has been associated with increases in individual behaviors that promote team performance (Briggs, Vreede & Reinig, 2003; Morris, Marshall & Rainer, 2002).

Trust, the willingness to be vulnerable to the actions of others (Mayer, Davis & Schoorman, 1995), and leadership, the ability to influence others, are two mechanisms used to promote positive team interactions and facilitate individual satisfaction (Carson, Tesluk & Marrone, 2007). There is rich and diverse literature on the benefits of trust in virtual teams but not leadership (Kayworth & Leidner 2001-2002; Robert, Dennis & Hung, 2009). Recent literature has indicated that virtual teams tend to engage in a type of shared leadership (Carte, Chidambaram & Becker, 2006). Shared leadership is the distribution of leadership among team members (Carson et al., 2007). Although virtual teams frequently use shared leadership we know very little about its implications in virtual teams (Balthazard, Waldman, Howell & Atwater, 2004; Carte et al., 2006). The few studies examining shared leadership in virtual teams have typically done it at the team level (Robert, 2012). However, shared leadership represents the contextual environment in which team members are embedded. Therefore, it is possible that shared leadership could both directly impact individual satisfaction and alter the impact of individual variables, such as trust, on individual satisfaction.

Copyright is held by the authors.
This paper takes a multi-level approach to understanding the impacts of shared leadership at the team level, and trust and satisfaction at the individual level. In doing so, this paper hopes to contribute to the literature examining shared leadership in virtual teams (Robert, 2012). The research question this study attempts to address is: “Does shared leadership alter the influence of individual trust on individual satisfaction?” This study has three goals: 1) to examine the impact of shared leadership on individual satisfaction in virtual teams. Recent work has examined the impact of shared leadership on individual satisfaction at the team level but little work has been done at the individual level (Robert, 2012). 2) To investigate the impact of individual trust on individual satisfaction. 3) To examine the interplay between shared leadership and individual trust on individual satisfaction in virtual teams. To address these gaps in the virtual team literature this paper proposes a multi-level research model that explains the relationship between shared leadership, individual trust, and individual satisfaction in virtual teams.

Related Literature and Hypotheses

Gibb (1954) originally suggested two forms of team leadership—distributed and focused. The nature of shared leadership bears the roles, responsibilities, and functions of leadership that are shared/distributed with two or more individuals within a team, rather than having the functions of leadership focused on a single individual (Carson et al. 2007; Lee, Lee & Seo, 2011). There are several terms that enable us to capture the concept of shared leadership across research: emergent leadership, collective leadership, and distributed leadership. Kocolowski (2010) comprehended these terms and defined the shared leadership, from his extensive literature review, as “a relational collaborative leadership process or phenomenon involving teams or groups that mutually influence one another and collectively share duties and responsibilities otherwise relegated to a single, central leader” (p. 24). Along with its broad conceptualization, research has identified different factors and dimensions of shared leadership. As a conceptual extension of shared leadership, emergent leadership was characterized by strategic goals, extensive networks, collaborative relationships, effective information processing, and focused action, by McIntyre (1999, p. 40). Another view is that shared leadership consisted of three dimensions: shared purpose, social support, and voice (Carson et al. 2007).

Shared leadership can foster an individual’s satisfaction. When leadership is shared, influence and power are distributed throughout the team. This distribution of power means everyone has a say in what actions are taken by the team. Shared leadership can also reduce conflict and promote intragroup trust and cohesion (Bergman, Rentsch, Small, Davenport & Bergman, 2012). A series of studies in organizations have found that employees are more satisfied with their job when leadership is shared (Loke, 2001). As such, we postulated the following hypothesis:

\textit{H1: Shared leadership will positively increase individual satisfaction.}

Along with shared leadership, trust is frequently associated with satisfaction. When individuals believe their teammates will act in their best interest they are less concerned about potential opportunistic behaviors (Robert et al., 2009). Team will feel more comfort and not have to expend effort monitoring the behavior of their teammates. This reduces the effort needed to accomplish work and facilitates positive team interactions. It has been empirically supported that trust of an individual team member is related to satisfaction. Jarvenpaa and Leidner (1999) has conducted a case study showing that satisfaction was higher when swift trust was formed in initial stage of collaboration and was maintained to the final stage. Hence, it is likely that as individual trust in the team increases so does individual satisfaction. Therefore, we can hypothesize:

\textit{H2: Individual trust will positively predict satisfaction of individual level.}

Multilevel theory implies that “higher-level units may shape or moderate relationships and processes in lower-level units” (Kozlowski & Klein, 2000, p. 8). We propose a cross-level moderation between shared leadership at the team level and trust at the individual level. The present study proposes that there is a substitutional relationship between shared leadership and trust on satisfaction. The ability to have a say in the actions taken by the team will be a strong and important predictor of individual satisfaction when individuals do not trust their teammates to make decisions on their behalf. However,
when individuals trust their teammates to make decisions on their behalf, having a say in what actions are taken by the team will be less important to individual satisfaction.

\[ \text{H3: There will be a moderating (substitutional) effect of shared leadership and trust on satisfaction.} \]

**Method**

**Participants**

Participants were enrolled at a large public university. Thirty percent of the students were females and ages ranged from 28 to 52 with a mean of 39. A total of 93 students, all US citizens, in 27 teams participated in the study. The size of the teams ranged from 3 to 5 with a mean of 3.7. The average team tenure was 47 days. Participants received extra credit for completing the survey.

**Data Collection**

Data was collected via an online survey administered to working professionals enrolled in a distance education program. The course instructors were not aware of the research question and none of the classes were taught by the author(s). The survey employed established multi-item 7 point Likert scales.

**Control Variables**

We used several control variables to reduce the possibility of alternative explanations. These included team size and team tenure, gender and racial diversity of the team.

**Independent Variables**

Team shared leadership was based on the density of shared leadership (Carson et al., 2007). Team members were asked to rate to what degree each team member displayed shared leadership. Density was calculated by dividing the average perceived leadership by 7, the total possible amount of leadership. The items measuring individual trust were taken from Simons and Peterson (2000). The items were “I know I can count on the other team members” and “I trust all of the other team members”.

**Dependent Variables**

Team satisfaction was measured using a scale that was a modified version of Briggs et al. (2003). Two of the items were “I was satisfied with how we completed the team project” and “Looking back I was pleased with how we completed the team project.”

**Results**

The virtual teams in this study were fully dispersed members taking classes in their home. These teams relied heavily on the use of electronic communication. Teams were asked about their technology use. Scales range from 0 to 6, with 0 indicating no use and 6 indicating extremely frequent use of electronic technology. The mean use and standard deviation of use per technology was: email 4.7 (.90), synchronous chat 1.9 (1.1), phone 2.04 (2.08) and video .2 (.47).

Psychometric properties were analyzed for each construct. The Cronbach’s alpha for individual trust and satisfaction was .95 and .90 respectively (Fornell & Larcker, 1981). Next, the intra-class correlation coefficient (ICC) was calculated for individual satisfaction. The ICC is used to justify using a multilevel analysis by providing evidence that team membership should be accounted for in the analysis (Bliese, 2000). ICC values of above .10 provide evidence that there is a significant team effect (Bliese, 2000). The ICC for individual satisfaction was .51, indicating a significant team effect.
A factor loading was done to assess convergent and discriminant validity. All items loaded at the .7 or above level on each of their constructs while no cross loadings were above .43. Both are clear indications of convergent and discriminant validity (Fornell & Larcker, 1981). The multi-level analysis was conducted with SPSS 20.0 mixed model package. The analysis, not shown because of space limitations, consisted of three models. Model 1 was the control model with size, tenure, gender and racial diversity predicted 3.75% of the variance in individual satisfaction. Model 2 was the main effect model predicted 71% of the variance in individual satisfaction. Model 3 included the moderation effect predicted 75% of the variance in individual satisfaction. The additional variance explain with the inclusion of the moderation effect of shared leadership and individual trust was significant ($F > .001$). The moderation effect was plotted (see Figure 1).

![Figure 1](image)

H1, shared leadership would increase individual satisfaction, was significant ($\beta = 3.4$, $p < .01$). H2, individual trust would increase individual satisfaction, was also significant ($\beta = .36$, $p < .001$). H3, the cross-level moderation effect of shared leadership and individual trust, was also significant ($\beta = -1.7$, $p < .05$). H1 and H2 were tested in model 2 while H3 was tested in model 3 which also included the main effects.

Discussion

This study provides several contributions to literature. First, this study contributes to the literature on satisfaction in virtual teams. Team interactions in virtual teams are often characterized as cold and impersonal (Chidambaram, 1996). As such, much research has been directed at promoting positive team interactions (Kayworth & Leidner, 2000). This study adds to that literature by highlighting that shared leadership also increases satisfaction at the individual level and not just the team level.

Second, this study contributes to the literature on leadership in virtual teams by examining the effects of shared leadership. One might have thought that shared leadership and trust might complement each other. Shared leadership would have stronger effects when coupled with high levels of trust in the team. In other words, allowing others to take the lead and act on your behalf is only satisfying when you trust others to act on your best interest. However, this was not the case. The results of this study suggest that the positive effects of shared leadership on individual satisfaction are due, in part, to everyone being able to have a say in what actions are taken by the team. As a result, when individuals trust their teammates to act on their behalf there is less of a need to have a say in everything and shared leadership has little influence on individual satisfaction. This suggests that the effects of both share leadership and trust reach a point of diminishing returns where any more has no additional effect.

Finally, this study contributes to the trust literature by examining the context-dependent effects of individual trust on individual satisfaction. Based on previous literature we would expect trust to always be an important predictor of satisfaction. However, allowing everyone to have a say in the actions taken by their team reduces the importance on trust on individual satisfaction. As a result, this study sheds light on
when and why trust increases individual satisfaction in virtual teams. In doing so, this study provides greater understanding of what mechanisms trust uses to influence satisfaction.

Limitations and Future research

This study is a cross-sectional study and like all cross-sectional research it is difficult to draw causal inferences. Future research should also investigate if shared leadership moderates the importance of trust for other variables. For example, trust has been known to facilitate knowledge sharing in teams. It would be interesting to see if shared leadership moderated the importance of trust on individual sharing of information. In essence, in teams with shared leadership, trust may not be that important to information sharing. Future research may also examine whether trust at the team level influences shared leadership. Prior research has already established that shared leadership in virtual teams influences trust (Lee et al., 2011). It is possible that trust at the team level could drive shared leadership. Future studies should test whether team trust increases the use of shared leadership in virtual teams.

Implications for Research

Shared leadership and trust can be viewed as coordination mechanisms. Teams can coordinate work by distributing leadership roles or by allowing an individual to take the lead. The results of this study suggest that both are equally effective at facilitating satisfaction. However, there are tradeoffs involved in both approaches. As the number of individuals involved in the decision making process increases so does the time and effort required to make decisions. But there are also benefits associated with having multiple individuals involved in the team decision making, such as attaining more views on a particular problem. Researchers may now want to study when and why which approach is more effective at facilitating performance.

References


Victoria L. Lemieux
iSchool
University of British Columbia
vlemieux@mail.ubc.ca

Abstract

Development of visual analytics technologies to support archival analysis, such as that conducted during arrangement and description, requires an understanding of the analytic task that the technology is being designed to support. Visualization technologies developed for archival application currently pay little attention to the analytic, cognitive and perceptual aspects of the design process. This paper, based on first results from a cognitive task analysis using verbal protocols, provides a broad brush description of archival analysis during arrangement and description. It then suggests some possible leverage points where visual analytics technology might be applied to support archival arrangement and description practices.

Keywords: archives, visualization, visual analytics, interactive visual interfaces

Introduction

Visual Analytics is being applied in many domains to assist analysts where there is a need to process masses of complex data and blend computational analysis with interactive visualization of the results (VisMaster, 2010). For this reason, it is an approach that is well-suited to the problems archivists face in arranging and describing vast quantities of digital records (Wash, 2012). What is Visual Analysis (VA)? VA is defined as the science of analytical reasoning facilitated by interactive visual interfaces (Thomas & Cook, 2005).

Development of visual analytics technologies requires an understanding of the analytic task that the technology is being designed to support (Fisher et al, 2011). The VA research cycle therefore includes working with decision-makers in the context of their organizations to characterize data and solutions in the situations in which analysis takes place. This work defines research questions for laboratory investigation, the results of which guide the design of new interactive visualization technologies and analytical methods that are evaluated in partnership with the decision-makers and their organizations.

Background Literature

Studies of the process of archival analysis in the context of arrangement and description are quite limited. Most of the literature on arrangement and description can be grouped into two categories. The first discusses the main concepts involved in arrangement and description (Duchein, 1986; Eastwood, 2000; Horsman, 2002), and the second tends to be normative or prescriptive in nature and deals with the practicalities archivists have to address when actually performing arrangement and description (i.e., by

Acknowledgements: The author would like to thank Lara Mancuso and Justin Chan for their research assistance. She would also like to thank the archivists who participated in the study, and the staff of the participating archival repository. Thanks must also go to the anonymous reviewers of this paper, whose comments have helped to improve it. Any errors or omissions remain the sole responsibility of the author. Finally, the author would like to acknowledge the generous financial support of Boeing Aerospace for this study.


Copyright is held by the author.
Only lately have archivists begun to explore the analytic processes involved in archival arrangement and description (Meehan, 2010; Yakel, 2003; MacNeil, 1995). Jennifer Meehan’s (2010) work comes closest to the objectives of this study. Meehan (2010) describes the process of archival arrangement and description as the archivist’s making sense of the records en route to contextualizing them. There is also technology development projects designed to support archival arrangement and description in a digital context, which have involved some study of users’ task requirements as part of the development process. The Archivotyperx Project at Simmons College uses web-based technology to teach students about the archival analysis process involved in arranging and describing archival records (Anderson et al. 2011). Hypatia is an initiative to create a Hydra application that supports the accessioning, arrangement and description, delivery and long-term preservation of born digital archival collections (AIMS, 2012). The tool includes plans for the development of a graphical user interface (GUI).

In terms of VA and archival arrangement and description, there are, at present, no studies that apply VA strictly to support archival arrangement and description, though there are a number of studies that use visualization to represent the outcome of archival analysis or which use VA to support access to archival collections and archival preservation decision-making. For example, an early and common visualization is a hierarchical node link graph, which is also typically used in representing organizational charts. These graphs have been used to represent the relationship between fonds (i.e., aggregations of archival records originating from the same creator or with the same provenance), sous-fonds (i.e., aggregations of archival records originating from an organizational sub-unit of the records’ creator), series (i.e., aggregations of archival records relating to a similar function), and items (i.e., individual documents) in an archival fonds (Library and Archives Canada, 2012). The introduction of XML encoded finding aids, particularly EAD, and wide spread implementation of descriptive standards such as ISAD(G) General International Standard Archival Description, have created opportunities for the introduction of new visualization tools that leverage archival descriptive metadata. Ian Andersen (Andersen & North, 2009) has explored the use of visualizations to represent multiple “dimensions” (e.g. relationships among persons, committees, etc.) in online finding aids. His project sought to visualize archival information by applying Ted Nelson’s ZigZag™ structure to two existing EAD finding aids. Robert Allen (2005) has explored the possibility of visually expressing hierarchies, networks, processes and timelines in US government archival holdings using Encoded Archival Description (EAD) to extract data and structure from source documents. Similar in purpose to Andersen’s work, Allen proposes an interactive approach to visualizing links between archival materials and has created a prototype interface that uses a “mass-spring model” to spatialize the relationship among the concepts (Provot, 1995). ArchivesZ, developed by Jeanne Kramer-Smyth (Kramer-Smyth et al. 2007), is a prototype of an information visualization tool that, in the same vein as Andersen and Allen’s work, leverages the structured data available in EAD encoded finding aids. By representing the distribution of subjects and time periods using the metric of total aggregate linear feet, Kramer-Smyth argues that ArchivesZ enables tool users to view total available research materials more quickly than they would by viewing a standard search result list. More recently, researchers at the University of Texas (Esteva et al. 2011) have developed a prototype VA system to aid archivists in identifying files requiring digital preservation and as a possible VA tool in support of archival research. They developed a visualization based on a space-filling treemap (Johnson & Schneiderman,1991) to present digital file-related metadata extracted from the collection at different levels of aggregation and abstraction. Mitchell Whitelaw (2009) has also developed a number of interactive visual interfaces to explore archival collections. Most recently, researchers at the University of Virginia (Nowviskie et al, 2012) have developed Neatline, a tool for geospatial visualization of archival collections.

**Methodology**

Efforts at applying visualization in the domain of archives have, so far, not placed great emphasis on the perceptual, cognitive and analytic aspects of the design process. Most efforts have made assumptions about the analytic requirements based on discussions with archivists, rather than rigorously researching requirements. In this study, we use cognitive task analysis and verbal protocols to observe archivists arranging and describing archival data (Trickett et al, 2007). The think aloud protocol is a method of gathering data where researchers ask participants to describe what they are doing as they perform a task. Two archivists were used in total for this study, who worked in a pair. This modified
approach to the use of the think aloud protocol, inspired by pair analytics (Arias-Hernandes et al 2011), overcomes one of the weaknesses of the think aloud method, which is that as individual participants become more cognitively engaged in their work, they articulate their thoughts less frequently. The research team found that working in pairs encouraged the archivists to continue to articulate and communicate their thoughts to each other throughout the process. Participants were asked to conduct their standard archival procedure on the records, while being observed using the think aloud cognitive task analysis method. Data gathering took place over five sessions. Sessions lasted between 80 and 100 minutes depending on the participants’ pace, and involved a full archival arrangement and description. Sessions were conducted in the participants’ normal archival location, so as to achieve a more naturalistic recording of the methods involved in archiving. Archival documents in the collection encompassed mostly analog images and files, but also contained some floppy disks and a film reel. Sensitive information within the fonds was removed prior to the arrangement. Archival material used in the study had not been previously arranged and described, so as to obtain a clear understanding of the entire process rather than repeating arrangement and description on a known set. This approach also allowed uncertainties to arise as normal during the process, since with previously processed material such problems may have been expected or already dealt with. During the data gathering sessions, participants were audio-recorded with their consent, and text transcriptions were made of each session. A brief interview also followed each session. In addition, observational notes on the environment or process were taken by the researcher-observers during arrangement and description of the collection.

Preliminary Findings

The research team observed that archival processing followed a three-stage sense-making process: in phase one, from first contact with the archival records, the archivists were concerned with gaining an overview of the structure of the archival fonds and creating a draft arrangement. In the second phase, the archivists confirmed and refined the arrangement structure and re-ordered archival documents, as necessary, and in the third phase, they described the final arrangement, documenting their description in an archival finding aid. One significant observation is that the initial process of analyzing the archival fonds and developing a mental model of its contents creates a significant cognitive load. Participants reported being much more fatigued in the first phase than they did in subsequent phases: the researchers observed them to be engaging in less joking during the process, and there appeared to be greater uncertainties. Part of this may be accounted for by the fact that the participants were still adjusting to being observed and performing the think aloud protocol, but observers remained silent throughout, and both subjects worked in their standard location using their own methods. More time was devoted to figuring out the timeline and context of the fonds; after this was determined, particularly in the second and third sessions, both participants were able to proceed more quickly. Further contextual information was easy to integrate once they had determined an overall mental model of the total arrangement, and fewer uncertainties were shown when new information became evident. The archivists were even able to infer the date and purpose of various files once this general mental framework was established. A more solid understanding of the fonds appears to have been established in the second session, when the participants began deciding on the series and their arrangement, despite not having completed arrangement on all of the files, and having focused mostly on rehousing archival documents. Even with these unknowns, participants’ final decisions on series closely mirror the ones they considered at this point. Similarly, initial predictions on the content of the individual folders or containers were more accurate once they had established more information about the creator of the files and had an initial mental model of the creator’s life. Any points of confusion that arose in the later sessions occurred because of inconsistencies in what they had established in their interpretation of the timeline and the context of the files they were working with. While neither archivist explicitly mentioned the development of a mental model, results of the experiment suggest that creating a mental model is a major part of the process of arrangement and description.

Another of the major findings is that the actual form of the archival documents is very important in conducting archival analysis. While this is not in itself surprising giving the well-known archival adage of “form following function”, the importance of form is of interest. The archival fonds with which the subjects worked was relatively small, comprising only three major divisions, but it still contained a large amount of information. It would have been impractical to have attempted to read every single document, or to explore every film reel or floppy disk. The large amount of content in even a small fonds means that trying
to create series using only content will run into difficulties, as this would require examining in depth every single file. However, both archivists were able to quickly examine and categorize hundreds of files into their various series by paying attention to form. For example, grant applications were very common throughout the files that the subjects received. After a fairly in-depth examination of a few of these grants, the archivists were able to quickly identify files that fit the visual description of a grant application, and were then quickly able to pick out relevant content, such as the size of the grant, the year, or the organization providing the grant. This then allowed the archivist to place the grant in a timeline of the author’s life, and to draw inferences as to what the records creator may have been doing with regard to their professional career, or even to ascertain the creator’s geographical location at the time.

**Design Implications**

Relevant information can be drawn from this initial study and applied to the design of a VA tool for working with digital archives. Identifying similarities in form is itself a sort of visual pattern matching. By quickly searching through visual representations of various digital files and identifying similarities in form, an archivist could quickly cluster files, and then rearrange them into suitable series, and outliers can quickly be brought to attention and categorized into an appropriate cluster, or form their own series. This potentially also allows for easier detection of duplicates, as well as cutting down the cognitive load for the archivists. Visual features could be supplemented by text mining for dates, which may assist in building a timeline, providing context for the archivist when trying to gain a sense of the fond as a whole as well as being of possible use to future researchers using that archive. The subjects also used a number of other programs or sources to help their process. Dictionaries and an online encyclopedia were used to define or elaborate on terms they were not familiar with, while physical implements like rulers were used to help mark their current location in a box as they arranged it, and using a text editor, they made note of the original organization of the archival data. In some cases, there were no available aids, such as when comparing two photographs to determine if they were duplicates, or trying to find whether an article had been published in a journal. Many of these aids could be integrated into a tool to assist the archivist’s visual analysis.

**Limitations and Conclusion**

The observations were done on only a single pair of archivists processing one collection and therefore it is uncertain to what extent the preliminary findings in this study are definitive or generalizable. Moreover, the study used physical archives, and it is still unknown to what extent the processing of entirely digital collections changes the analytic process. Nevertheless, it is still possible to use the findings to produce a prototype VA system for archival arrangement and description which can then be evaluated and field tested by users to refine the design. The research team is now working on the development of the prototype. In addition, once the prototype is developed, the resulting VA tool can be used as a test-bed to study archivists’ interaction with the tool as they arrange and describe digital records in order to yield additional insights into the analytic processes involved in archival arrangement and description. These insights then can be used to support refinement of the model resulting from this research and generalizability of the research findings. The researchers also hope to work with other researchers involved in projects supporting archival arrangement and description in digital contexts (e.g. Archivopteryx and Hypatia) to expand the study of the analytic processes involved in archival arrangement and description across a larger group of study participants and range of archival collections.

**References**


470


Whitelaw, M. Visualising Archival Collections: The Visible Archive Project. *Archives and Manuscripts* 37, 2, 22-40.

Using Vignettes to Study Conflict Management Practices in Enterprise Architecture Initiatives

Guangxuan Zhang  
Penn State University  
gxzhang@psu.edu

Sandeep Purao  
Penn State University  
spurao@ist.psu.edu

Abstract

This note presents an ongoing effort to study conflict management practices in EA initiatives. We suggest an adaptation of the vignette technique to collect data and describe how it was used as part of an exploratory study. The analysis so far reveals four conflict management practices – avoiding, dominating, compromising, and transforming. We describe situations in which the four practices were used, investigate some of the intricacies revealed by the vignettes and elaborate on the potential for using vignettes to study conflict management practices.

Keywords: conflict management, enterprise architecture, vignette

Introduction

Conflict is pervasive in information system (IS) projects (Barki & Hartwick, 2001), appearing at the intrapersonal (Ruthner, Hardgrave, & McKnight, 2008), interpersonal (Barki & Hartwick, 2001; Johnson & Cooper, 2009; Sarker & Valacich, 2010), and intergroup levels (Kumar & Dissel, 1996). According to Meissonier and Houzé (2010), conflict is caused by different understandings of technology, technology-enabled structures, and power changes; different working styles; and interpersonal dissonance. Although organizational literature shows that conflict can positively influence performance (Baron, 1991; Janssen, Van De Vliert, & Veenstra, 1999; Jehn, 1995; Nemeth, Connell, Rogers, & Brown, 2001), conflict has negative impacts on individuals and leads to detrimental outcomes (Barki & Hartwick, 2001; Sarker & Valacich, 2010).

Understanding and managing conflict is, therefore, an important concern for organizations that embark on Enterprise Architecture (EA) initiatives. Organizations must engage a variety of stakeholders to (re-)define current and target goals, and chart a path to transition between the two (Jonkers et al., 2006; Lankhorst, 2009; Lin & Dyck, 2010; Schmidt & Buxmann, 2011; Winter & Schelp, 2008). These actions often entail significant organizational change and require re-alignment between business and IT functions (EARF, 2009; Gartner, 2008; Schmidt & Buxmann, 2011). All of these actions make conflict an inherent outcome. The very nature of EA requires input from participants that include IT departments, top managers, system users, and external participants such as consultants and vendors (Jonkers et al., 2006). The sheer scope of EA initiatives and the diversity of interests they entail means that EA initiatives are likely to produce more conflict among participants compared to traditional IS projects (Donaldson & Preston, 1995; Winter & Schelp, 2008). Furthermore, understanding and managing conflict doesn’t simply prevent dysfunctional consequences, it can also leverage the positive creativity unleashed in conflict situations. Both are critical prerequisites for realizing the benefits of EA initiatives (Smolander & Rossi, 2008; Van Steenbergen & Brinkkemper, 2009).

We describe here an ongoing research effort to investigate conflict and conflict management practices in EA initiatives using a novel extension and application of Vignettes (Miles, 1990) to collect and analyze data from a service provider in the Federal Government sector, and one of its client agencies. Our adaptation allows collection and analysis of qualitative data that overcomes some of the shortcomings of other research methods. Our data analysis and experience using the technique highlights two findings: 1) initial analyses indicate four conflict management practices used in EA initiatives.


Copyright is held by the authors.
initiatives; and 2) we reflect on the use of an adaption of the Vignettes technique as an efficient approach for collecting data in organizational contexts.

Prior Research

Workplace conflict has attracted much attention in the fields of organizational science and psychology. These scholars define conflict as an interactive process beginning with perceived incompatibility between interdependent social entities that unfolds to generate undesirable consequences (De Dreu & Gelfand, 2008; Putnam & Poole, 1987; Thomas, 1992). Scholars have described the causes (Kriesberg, 2007; Mikolic, Parker, & Pruitt, 1997; Walton & Dutton, 1969), process (Druckman, 2001; Pondy, 1967; Pruitt & Kim, 2004; Thomas, 1992), and outcomes of conflict (De Dreu, 2006; Jehn, 1995; Nemeth, 1986; Simons & Peterson, 2000). Prior work has described models for conflict management (see summary in (Rahim, 2010)), which have since been widely accepted (Olekalns, Putnam, Weingart, & Metcalf, 2008; Rahim, 2010; Thomas, 1992), and use the axis of concerns for self versus others to suggest five patterns: collaboration, yielding, contention, avoidance, and compromise.

Much of this work has been applied in the context of IS development (Smith & McKeen, 1992). But, EA initiatives—a superset of IS projects (Barki & Hartwick, 2001)—represent a fertile arena for studying conflict and conflict management. Like IS projects, EA initiatives involve business-IT alignment, and require close collaboration between IT and other departments (Jonkers et al., 2006), and conflict between participants is almost inevitable (Donaldson & Preston, 1995; Winter & Schelp, 2008). To our knowledge, no study has investigated conflict and conflict management practices in EA initiatives, with the exception of Meissonier and Houzé (2010) who report ‘avoidance’ as an effective strategy in initiatives similar to EA. This remains an area in need of deeper investigation.

Methodology

A Novel Research Methodology with Vignettes

Given the paucity of research, we believe a qualitative approach is more appropriate for research this domain. Although a quantitative approach might be tempting, it would likely strip the essential context of EA projects. Data collection might allow us to identify conditions and outcomes (Baron, 1989; De Dreu, 2006; Jehn, 1994; Nemeth, 1986; Putnam, 1994), but cannot shed light on the behavior, perspectives, and historical elements that also play roles. Self-reporting instruments might be better, but still fall short in capturing and describing behaviors (Olekalns et al., 2008).

The nature of conflict as the “dark side” of the workplace (Raver & Barling, 2008) is an obstacle to collecting qualitative data (Kolb & Putnam, 1992). Data collection might be hindered through a reluctance to share data, and informants may respond in a defensive or biased manner. Jehn (1997) suggests collecting data by asking informants to “describe a conflict a typical employee in your team would experience.” While useful as a prompt, in the extreme (i.e. rejecting authentic narrative) it may have the unintended consequence of foregoing rich detail. Case studies (Markus, 1983; Orlikowski, 1993; Yin, 2009) and ethnographic studies (Myers, 1999) overcome this concern with techniques to ensure anonymity and encourage participation. But, the cost and effort can be prohibitive. It can also be difficult to convince organizational participants to engage in data collection requiring significant effort without the promise of immediate benefits.

To overcome this mix of concerns, we developed a novel data collection technique that adapts Vignettes as the key unit of analysis to increase the amount and quality of data. Miles (1990) describes a Vignette as a snapshot or a ‘mini-movie’ of a practitioner at work. It reflects a recent episode of practice in two dimensions: describing the practice, and producing thoughtful explanation. The event isn’t simply reported, participants are also encouraged to reflect on their own engagement in it, producing a rich dense account. Researchers have used vignettes as “icebreakers” (Barter & Renold, 1999) for surveys (Alexander & Becker, 1978) and interviews (Schoenberg & Ravdal, 2000). Our proposal is to use Vignettes as a “first-class unit” as part of a systematic, structured approach to collect and organize qualitative data (Miles, 1990). The context is especially relevant because Vignettes allow not only the benefits described above, but also provide a platform from which observations from different perspectives can be quickly integrated before they are subjected to member-checking.

The process begins with brief, intensive interviews with informants. The researchers then generate multiple Vignettes, each describing a situation of interest. When later shared with the
informants, these vignettes may open up the informant’s thinking while at the same time describing their behavior. Researchers and informants then reflect on and further enrich the narrative. An ancillary benefit is that vignettes can be shared rapidly with informants to demonstrate how “anonymizing” them protects confidentiality. This in turn encourages participants to be more open (Alexander & Becker, 1978), eliciting more candid accounts (Burstin et al., 1980). In contrast to case studies, which can be time-consuming and require high data inquiry skills (Myers, 2008), Vignettes require less time and effort (Miles, 1990); narrow in scope, informants can directly develop and refine them via member-checking (Schoenberg & Ravdal, 2000). As a set of multiple data points, vignettes provide the foundation for data and pattern analysis.

Research Design and Process

This project aims to illuminate conflict management practices in EA initiatives. This note reports outcomes of the first step in the project, the exploratory study. We began by cultivating a connection with SSP (pseudonym), a shared service provider for U.S. federal agencies. SSP offers an integrated service model that includes infrastructure hosting, applications management, technical support, user support, transaction processing, and business services. SSP takes pride in its use of EA-thinking and makes a conscious effort toward integrating an EA-perspective in their work with clients. GOV (pseudonym) is one of their infrastructure hosting service customers, that is a federal agency with about 50,000 employees working in headquarters and regional offices. Our interaction with SSP and GOV was guided by the research methodology outlined above, with data collection, formalization of vignettes, and data analysis. Figure 1 outlines the steps.

Figure 1. Research Design and Progress

In Phase 1, the research team interviewed informants from SSP and GOV. Multiple sources of data, including interview notes, presentation slides, and project documents were collected. Based on these materials, the research team applied the process suggested by Angelides and Gibbs (2006) to identify eight vignettes, each representing a conflict scenario.

In Phase 2, the vignettes were sent back to two representatives (one each from SSP and GOV) for member-checking. Based on feedback, two were removed because they overlapped with other vignettes, and changes were incorporated to refine the narrative in several others. The resulting six vignettes now represented a non-trivial, rich description of various conflict situations, including descriptions of how they were managed by different stakeholders.

During Phase 3, the vignettes were put through two levels of analysis—content analysis and pattern matching (Myers, 2008; Yin, 2009). First, open coding was applied (see Figure 2).
Vignette 2

In 2009, GOV decided to try a pilot with a social media implementation to see whether it could improve communication and cooperation within and across departments. They selected IBM Connections as the software and decided to host it on SSP's server. When they started to negotiate on the service terms, the **top manager of GOV** required that the system must be launched in one month. The **CTO of SSP** rebuffed that one month was unreasonable because more time were needed to install the new software and test its compatibility to the existing components. However, the manager **did not accept any compromise**, joking that he would give one more day to **SSP**. At the end, **SSP** gave in and agreed on putting the services online in one month. The system was actually launched in time.

<table>
<thead>
<tr>
<th>Conflict parties: Top manager of GOV and CTO of SSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power distance between conflict parties: <strong>High</strong></td>
</tr>
<tr>
<td>Conflict issues: project timeline</td>
</tr>
<tr>
<td>Is the issue related to business structure change? <strong>No</strong></td>
</tr>
<tr>
<td>Does the issue require strategic concern? <strong>No</strong></td>
</tr>
</tbody>
</table>

Table 1
Open Coding Summary

<table>
<thead>
<tr>
<th>Vignette</th>
<th>Parties</th>
<th>Issues</th>
<th>Structured Change</th>
<th>Power Distance</th>
<th>Strategic Concern</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1: System platform</td>
<td>Business Department / IT Department</td>
<td>Technology</td>
<td>No</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>V2: Project schedule</td>
<td>Manager / Vendor</td>
<td>Project Management</td>
<td>No</td>
<td>High</td>
<td>No</td>
</tr>
<tr>
<td>V3: Selecting software applications</td>
<td>Manager / Employees</td>
<td>Technology</td>
<td>No</td>
<td>High</td>
<td>No</td>
</tr>
<tr>
<td>V4: Resisting new framework</td>
<td>Vendor / IT Department</td>
<td>Technology</td>
<td>Yes</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>V5: Cutting project to budget</td>
<td>IT Department / Manager</td>
<td>Project Management</td>
<td>No</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>V6: Resisting new technology</td>
<td>Manager / Vendor / IT Department</td>
<td>Technology</td>
<td>Yes</td>
<td>High</td>
<td>No</td>
</tr>
</tbody>
</table>

While it may be argued that open coding requires participant-generated text, we were careful to use phrases and terminology directly provided by respondents—further reinforced through the member-checking phase—resulting in refinements expressly suggested by participants. Thus, the vignette text was a direct reflection of participants’ intent and ideas, making content analysis an appropriate technique to employ. Table 1 shows the cross-section of situations that lead to conflict when GOV partnered with SSP.

Content analysis was completed through pattern matching (Yin, 2009) to distinguish conflict management practices. The team completed this phase by returning to the informants again, and using prior research as the basis for comparison and extension, identified multiple dimensions for each vignette, including: structural change, power distance, and strategic concerns. Each reflected a key element that has been investigated in prior work (Pruitt & Kim, 2004; Simons & Peterson, 2000; Thomas, 1976).
Results and Discussion

Early Results: Conflict Management Practices

The vignettes revealed four conflict management practices (see Table 2):

1: **Avoidance** Often a temporary solution (Kerzner, 2009), avoidance postpones conflict with the hope that a resolution would emerge before the situation became critical or that the conflict would "resolve itself." In Vignette 4 the IT department declined a vendor's proposal by withdrawing. In Vignette 6, the IT department refused to migrate to shared service by postponing for "evaluation."

2: **Domination** Here, the party with greater power makes decisions, forcing others to follow (Pruitt & Kim, 2004). When faced with different project timelines in Vignette 2, GOV's manager set the project deadline and forced SSP to buy in. In Vignette 3, the office manager imposed his preferences on subordinates. Domination was not used in Vignettes 5 and 6, possibly because the issues were significant to the organizational structure, or related to the business model or strategic concerns.

3: **Compromise** Compromise seeks the middle-ground position. Conflict parties make concessions to reach a mutually acceptable decision (Rahim & Bonoma, 1979). Vignette 5 showed a downsized project finally accepted when it matched management's budget concerns with IT's enthusiasm.

4: **Transformation** Transforms the conflict into new factors that may be relevant but not essential to the original conflict; suggests that conflict happens within organizational context, manifesting and resolving through social and organizational factors. Vignette 1 showed conflict rooted in different preferences about software platforms resolved via cost consideration.

Table 2

<table>
<thead>
<tr>
<th>Conflict Management Practices and Associated Vignettes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>Avoidance</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Domination</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Compromise</td>
</tr>
<tr>
<td>Transformation</td>
</tr>
</tbody>
</table>

Discussion

Although this exploratory study developed limited vignettes, the potential for further analysis and interpretation is clear from the data. For example, Avoidance was observed where the conflict had the potential to change the structure of the IT department, which chose to avoid the conflict to prevent a loss of power (Markus, 1983). Domination was observed in situations with unequal power positions (Pondy, 1967) or unbalanced negotiation power (Pruitt & Kim, 2004). Compromise appeared where stakeholders had widely different concerns and goals, but the conflict resolution effort focused on ensuring that the interest of every stakeholder was incorporated (Jonkers et al., 2006). Transformation was observed when the conflict issue appeared to be substantive but not critical to all conflict parties. This assessment helped identify other concerns, which then cut down the negotiation time to arrive at a resolution. As our analysis proceeds, we are deepening our interpretation beyond the initial observations shared here. For example, we are trying to understand how these vignettes can provide a window onto the viewpoints of different stakeholders; whether we can place the situations described in each vignette at different temporal stages
of an EA initiative; and, whether further analyses (such as multiple practices) may be discerned from the situations described in each vignette. An important component of this next step is to map interpretations against results from prior work much like the step suggested by Eisenhardt (1989).

Concluding Remarks and Next Steps

We have described an ongoing research effort that employs vignettes to investigate conflict management practices in EA initiatives. Our effort demonstrates that vignettes provide a feasible approach to data collection and units for analysis. And, this first round of analysis has identified four conflict management practices: avoidance, domination, compromise, and transformation. An initial mapping against the situations where each practice is used is also indicated. Our work continues to deepen the analysis in this exploratory study, and to collect further data to enrich our data set.

References


Information Flows in Events of Political Unrest

Karine Nahon, Jeff Hemsley, Robert M. Mason, Shawn Walker, Josef Eckert  

University of Washington

Abstract

Social media, as the set of tools typified by blogs and other social networking platforms, is creating a user-generated dynamic, complex information ecosystem. The flow of information across multiple platforms means that traditional media gatekeepers (newspapers and other ‘mainstream media’) become just one of many pathways by which we learn about and make sense of new information. This research note reports preliminary results of a study based on a dataset of more than 65 million tweets related to the Occupy Wall Street movement, coupled with searches of LexisNexis, to examine information about six events related to Occupy sites in Maine, New York City, Oakland, and the University of California - Davis. The study seeks to understand the relationships among newspapers, blogs, and Twitter as users of each platform report and comment on these events. The preliminary results suggest that the platforms perform distinct but overlapping roles at different periods in the information diffusion life cycle.

Keywords: information flows, social network, Occupy, Twitter, social movements

Introduction

Recent political unrest and social movements are characterized by multiple, diverse practices, sites of protests, and are set in varied contexts (e.g., the revolutions frequently termed “The Arab Spring”; the Indiginados movement in Spain; the tents movement in Israel; and the Occupy movement in the United States). One unifying element is the growing use of social media and ways in which social media and traditional media interact. This research compares information flows in the mainstream media, Twitter, and the blogosphere using six cases, each an event related to the Occupy Movement in 2011 and 2012, and each fitting into one of three event types: 1. Serendipitous events, not planned by either the protesters or authorities; 2. Events planned by authorities; 3. Events planned by the protesters. We posit that the type of event (planned or unplanned) can create distinctive information flow patterns among the three platforms.

This initial work is part of a broader project that addresses questions such as how information gains the attention of publics (e.g., through traditional media, alternative media like blogs, or an emergent mix of these in conjunction with social networks). In an interconnected world, we expect a mix of information flows and sharing practices among these platforms. In such an ecosystem, what is the interplay between and among these platforms? What kinds of events influence where information will first emerge? Is there a connection between the type of event and the way information is distributed? Answering these questions will help us to further refine our understanding of how communities create meaning from information and consequently influence attitudes.

Acknowledgement: NSF, award number 1243170
Copyright is held by the authors.
Information Distribution and Social Movements

The pervasive use of social networking sites has provided an opportunity for researchers to study information flows and the roles of information platforms during political unrest (Gaffney, 2010; Lotan, Graeff, Ananny, Gaffney & Pearce, 2011). An increasing number of researchers are examining information flows in networks. Twitter and blogs are associated with rapid information flows across information-sharing networks, and well-positioned influencers can shape to a certain degree how information flows (Hemsley & Mason, 2012; Lotan et al., 2011; Nahon, Hemsley, Walker & Hussain, 2011). Gomez-Rodriguez et al. (2012) trace paths of diffusion and influence through networks over which contagions propagate, and Leskovec et al. (2007) describe the distribution of information forwarding chain topologies in the blogosphere. Nahon and Hemsley (2011) examine the factors that influence blogs to co-link to videos, thus driving information diffusion.

In this work we seek to add to the nascent body of literature that focuses on information flows in the context of collective action or social movements. Lotan et al. (2011) studied dissemination of news via Twitter during the Tunisian and Egyptian revolutions in 2011 and identified six types of roles that impact the practices of information distribution. Constanza-Chock (2012) examines media practices during social movement activities, focusing on the Occupy movement and Mazone (2011) provides a visual comparison of newspaper coverage of the Occupy movement with its presence on social media sites.

This research note presents preliminary results from a study that examines the life cycle of information sharing about selected “Occupy” events through Twitter, blogs, and mainstream media (printed and online). Our research addresses the following questions: in a time of unrest, how is information about a given event distributed via these channels? Does one platform tend to distribute information earlier than others? Can we see evidence of information moving from one platform to another? We address these questions by examining six specific events.

Data Collection

Twitter data are drawn from the corpus of 65 million tweets that the Social Media Lab at the University of Washington1 has been collecting since October 2011. The data were collected using over 330 Occupy related search terms and Twitter’s Streaming API. Mainstream media and the blogosphere data were drawn from LexisNexis.

We selected six events to compare as cases, each of which fits into one of our three event types: 1. Serendipitous events, not planned by either the protesters or authorities; 2. Events planned by authorities; 3. Events planned by the protesters. We posit that the type of event (planned or unplanned) will affect the information flow among the three platforms (Twitter, mainstream media and blogs). See table 1 for description of these events.

1 http://SoMeLab.net
Table 1

<table>
<thead>
<tr>
<th>Type of Event</th>
<th>Name of Event</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serendipitous</td>
<td>Bomb in Occupy Maine camp</td>
<td>10.23.2011</td>
<td>A bomb was thrown early in the morning at the Occupy Maine camp</td>
</tr>
<tr>
<td>Serendipitous</td>
<td>Scott Olsen injury</td>
<td>10.25.2011</td>
<td>During the clearing of the Occupy Oakland camp, Scott Olsen, a veteran, was hit by a police tear gas canister and injured</td>
</tr>
<tr>
<td>Serendipitous</td>
<td>UC Davis pepper spray</td>
<td>11.18.2011</td>
<td>University police pepper spray a group of seated demonstrators during an Occupy movement demonstration</td>
</tr>
<tr>
<td>Planned by protesters</td>
<td>Occupy Oakland port shutdown</td>
<td>2.11.2011</td>
<td>A demonstration event planned by the Occupy Movement to shut down the Port of Oakland</td>
</tr>
<tr>
<td>Planned by protesters</td>
<td>Marking six months to the Occupy movement</td>
<td>3.17.2012</td>
<td>Occupiers attempted to re-occupy the Zuccotti Park to mark the six month anniversary of the Occupy movement</td>
</tr>
<tr>
<td>Planned by authorities</td>
<td>Clearing the LA Occupy camp</td>
<td>11.30.2011</td>
<td>Police moved to clear the LA Occupy camp just after midnight</td>
</tr>
</tbody>
</table>

The Approach

We used an iterative approach to extract the data from Twitter (our collection) and from LexisNexis news and the blogosphere. We first identified keywords we thought would produce the best search queries for each event. To refine our search terms, we selected two days of tweets after a given event and conducted a preliminary manual content analysis, identifying the most common event related words used in tweets. We used the same keywords for our final queries to extract data from the SoMe Lab Twitter archive and LexisNexis. We bounded all six events temporally to 48 hours before and 10 days after the event.

Precision (estimating Noise). The Twitter data included hundreds of thousands of tweets, many of which we consider noise, unrelated to the event. To estimate the volume of noise we randomly sampled 1000 tweets from each event-dataset and manually coded each tweet as related to the event (related) or unrelated (noise). After removing the noise we compared the remaining Twitter data samples with the news and blogging samples. We posited that the random sample had the same temporal pattern (number of tweets per hour) as the actual data. We then produced synchronized plots for each event to permit a visual comparison. Future work will include more cases, larger samples and a time-series analysis.

Preliminary Findings

Figure 1 (below) plots the information flow of posting information related to the event in Twitter, blogosphere and mainstream media over a period of twelve days.

Findings from figure 1:
1. First to appear: it is clear that for serendipitous events and events planned by protesters Twitter activity appear before the mainstream media (online and printed) and blogs. However, for the event planned by the authorities, the mainstream media discussed the event well before it happened.
2. Life Cycle: The life cycle of information distribution looks similar in all three cases when events are unplanned (serendipitous). However, planned events demonstrate different patterns. For events planned by authorities, tweet volume quickly ramps up to a spike and then decays and goes flat as interest in the topic fades away, whereas mainstream media and blog volume occurs before, during, and after the event, though somewhat unevenly. In events planned by protesters, the life cycle is similar across the media platforms but news and blogs lag behind Twitter.
3. Serendipitous events: From figure 1, we see differences in two dimensions: a) the volume of posts in blogs and the mainstream media is higher in serendipitous events than in planned events; and b) there is a lag between the event and the beginning of the information sharing about the event for all three media platforms.
4. **Planned events**: planned events exhibit different information practices than serendipitous events. Most of the Twitter activity happens in real-time, at the time of the event. In events planned by protesters, Twitter volume appears differently than in the news and blogosphere. Users in Twitter lose interest quickly (about 2 days after the event). In contrast, we see that the blogosphere and mainstream media regenerate a conversation about the event a few days after the event. Preliminary content analysis suggests that Twitter serves as a vehicle to transfer practical information about what is happening in near-real time.

![Figure 1. Frequencies of Information Sharing on Twitter](image)

Figure 1. Frequencies of Information Sharing on Twitter
This preliminary content analysis of the sample Twitter data suggests other observations and distinctions that may be worth pursuing in later studies. For a planned event (e.g., clearing of the LA Occupy camp or the general strike/attempt to close the Port of Oakland), the anticipation of the event becomes part of the news cycle, regardless of the platform, with the expectations, results, and analysis shared in what might be a predictable cycle.

The stories of the unplanned (serendipitous) events, however (such as the injury to Scott Olsen from a tear gas canister, or the pepper spraying of the seated students at UCDavis), do not have the build-up of anticipation—they exhibit a sharp peak. The dropoff can be rapid (as in a normal news cycle), but the reactions to at least these two events maintain a higher level of interest. These two events become more than a news item that was soon forgotten. They became the focus for additional discussions and discourse on the different platforms; they became *memes*, serving as a kind of shorthand and boundary spanning event that engaged a wide range of participants. The engagement kept the original story alive in a different form, as, for example, when Megyn Kelly on Fox news observed that pepper spray is really just a vegetable product (2011), leading to discussions (often humorous) that perpetuated the discourse.

**Discussion**

The patterns in Figure 1 suggest that information distribution practices may be associated with the types of events and type of media platform. For example, the mainstream media discusses events planned by authorities well ahead of time. This may reflect relationships between authorities and media outlets where authorities call a press conference to indicate how they plan to deal with protesters. This gives them an opportunity to signal their intentions and frame the situation beforehand. Our preliminary analysis indicates that Tweeterers and bloggers note this signaling and framing; they feel excluded from what they see as a close relationship between authorities and the mainstream media. We also noted a higher volume of information about serendipitous events than for planned events. This could reflect our case selection, but it could also be evidence of interaction between social media and the mainstream media: if an event draws a great deal of attention in social media, it becomes newsworthy and news stories are generated. These news stories in turn gain attention from social media users, who then discuss not just the event, but the mainstream media’s response to it. We find some evidence of this in our reading of tweets and plan to explore this more in future work. Finally, the question of lag in serendipitous events between the occurrence of the event and the beginning of the life cycle of the information flow can be attributed to the fact that users rely on gatekeepers. Once these gatekeepers (journalists or bloggers or Twittereres) post information about this event, then the public starts to share information.

**Limitations**

The work reported here is preliminary and we cannot draw general conclusions from our findings. However, the effort has provided us with the basis for improving our research approach and with the basis for focusing our research questions.

We find that when examining Twitter and other social media data, we must carefully define the scope of our search. Social media is extraordinarily rich; from the wide range of contributors to the universe of data available, researchers have considerable flexibility in conducting studies. The resulting dataset that we develop for our analyses is extremely sensitive to the search parameters. To achieve coverage we’d like to have in order to draw robust conclusions, we are likely to get a lot of noise in the form of irrelevant tweets or posts. On the other hand, narrowing the scope to achieve the *precision* (reducing the noise) means that we may miss critical posts or tweets.
Bibliography


E-books in Academia: Expectations and Challenges

Irene Lopatovska
ilopatov@pratt.edu

M. Cristina Pattuelli
mpattuel@pratt.edu

Leanora Lange
llange@pratt.edu

Victoria Ludas Orlofsky
vludas@pratt.edu

Pratt Institute
School of Information and Library Science

Abstract
This paper reports on the results of a pilot study conducted as part of a larger project aimed at understanding the evolving role of e-books in academia. The study used a survey, a scenario-based usability test, and interviews to collect data on e-book reading habits, specific user expectations, and issues related to the usability of an academic library interface for e-book access. Future work will investigate factors affecting user preferences and patterns of use in context, including routine use of academic e-books.

Keywords: e-book, e-reader, reading habits, technology adoption, academic libraries

Introduction
Every day academic libraries make significant investments in e-book packages or individual e-book titles with little or no information on how students and faculty use e-books and e-readers in the course of their work. This paper describes the initial phase of an ongoing research project conducted in partnership with several academic libraries on the adoption of e-books and e-readers in academia. The study described here was conducted in collaboration with a major academic library. Factors investigated included users’ familiarity with and expectations toward academic e-books, and the impact different interfaces and devices have on users’ research and reading habits. The project aims to provide insight and recommendations that would inform library decisions in the areas of collection development, instructional programming, and library catalog interface design. We ultimately hope to gain a better understanding of the evolving role of digital media in education and research within the changing landscape of scholarly reading practices.

Relevant Literature
Though issues concerning e-books in academic libraries have been discussed since before 2000, there has been a marked increase in research conducted on the subject in the past decade as the medium and related technologies have evolved. The bulk of the work has fallen into a few main categories, including adoption, usability, vendor relations and offerings, and collection development.

A large number of studies have focused on issues related to the acceptance and use of e-books. Most of the studies found that while awareness of e-books has increased, usage has yet to catch up, and preference for print continues to be strong among students as well as faculty and staff (Woody et al., 2010; Croft & Davis 2010; Li et al., 2011). This trend appears to be changing with the introduction of more sophisticated e-readers, but while students with e-readers show an increasing fondness for e-books, they

Acknowledgements
We would like to thank Aimee Slater for assisting with data collection and analysis.


Copyright is held by the authors.
remain a relatively small population (McCarthy 2011; Pearson 2011 and 2012). Many studies have shown that the most heavily used e-books in academia are in the areas of computer science, engineering, business, economics, management, and social sciences (Nicholas et al. 2008; Abdullah & Gibb, 2008; Simon, 2011). Though e-book use in the arts and humanities has lagged, for the most part this appears to be more related to the availability of e-books in those fields rather than the preference of arts and humanities scholars. While print is still preferred in cases of extended reading, academics in the arts and humanities are both aware of and open to the medium overall (Fischer et al. 2012).

Another set of topics frequently discussed in the literature relate to the evaluation and comparison of vendor products and services. No vendor interface has proven to be particularly popular with patrons (Kimball et al., 2010; Lippincott et al., 2012). Many of the negatives cited include problems with the interface functionality and navigation of e-book reading programs and the discomfort of reading long passages on a screen.

This examination of the literature indicates that while e-books are gaining popularity among academic users, there are a number of issues that still hinder their wide acceptance. In an effort to better understand the use of e-books in academic environments, the initial phase of our study focused on investigation of the user expectations and the challenges pertaining to the use of e-books for scholarly pursuits.

**Methods**

The data were collected in the spring of 2012. A pool of four faculty and eleven students from the partner institution representing diverse demographic characteristics (e.g., age, sex, race) participated in the study. Participants were selected from a pool of volunteers who responded to the recruitment flyers posted around the campus. The study used a purposive sample of faculty and students who had experience using academic e-books. The study site was a usability lab located on the Pratt Institute’s Manhattan campus. Each session involved three stages: first, participants were asked to fill out a short survey that collected demographic data and data on their e-book and e-reader use habits; next, participants were asked to complete several search tasks that required them to access e-books through the website of the partner library and also through other website(s) of their choice (e.g. Amazon). The tasks were designed to investigate specific features of the library interface for accessing e-books and solicit participants’ suggestions for improving the site’s usability. Finally, participants were interviewed about their search experience in a post-task debriefing interview. Interviews lasted an average of 20 minutes. The interview data were audio-recorded and captured in the researchers’ notes.

**Data Analysis and Discussion of Results**

Survey responses of the fifteen participants were summarized using descriptive statistics. Ten out of the fifteen participants reported owning an e-reader. The most popular reason for owning the e-reader was convenience. Other frequently cited reasons were cost, text searchability, and interactive features. The most popular channel for obtaining academic e-books was identified as the academic library. Participants cited personal purchases and public/non-academic libraries as other channels they favored. Twelve participants reported using the partner institution’s e-books for research, while three participants reported never having used e-books from the partner institution in the past. Figure 1 summarizes survey findings related to the frequency of e-book and e-reader use as well as the reported experience with using the institution’s e-books and e-books acquired from other sources. We collected data on the use of e-readers in general, without differentiating between their academic and non-academic uses.
Results revealed several trends in participants’ habits of using e-books as well as in their use of e-readers:

1. While 6 participants reported using e-readers daily or few times a week, only 2 of these frequent uses were associated with academic e-books. Three participants have never used e-books for academic purposes, while one participant reported using academic e-books on average once a month and 4 participants reported using e-books for academic purposes several times a year. No one used academic e-books on a daily basis.

2. The strongest motivation for using e-books was the instant access to content that made it unnecessary to visit the library in person, followed by the unavailability of a printed version of the same book.

3. Most participants reported having an easy and intuitive experience while using e-books; fewer participants reported having positive experiences associated with the use of e-books from the partner institution’s library compared to the books accessed through other channels.

While results from the search task portion of the study were specifically focused on technical usability issues related to the partner institution’s library interface and would be less relevant to report in the context of this paper, data from post-task debriefing interviews offered insight into user expectations and issues related to academic e-book usability. Questions focused on the motivations behind the search strategies participants adopted to find specific academic e-books as well as on their level of satisfaction with interface design issues (e.g., wording of labels). More general questions solicited participants’ accounts of their experience using e-books in the context of their academic institution and, when applicable, using their personal e-reader. Interview data were coded and grouped in categories, and six major themes emerged from the content analysis. Responses related to more than one category per question were copied under each relevant category. Figure 2 offers a visualization of these themes and their relationships arranged across a continuum of issues of familiarity and issues of access.
Figure 2. Model of themes and their relationships as identified in post-task interview data.

The theme at the center of the analysis was users’ desire for a simple, easy interface. Of the fifteen participants, 11 proposed adding more detail to the labels of the links to make it easier to navigate across them. Some offered specific suggestions for a more user-friendly interface such as adding “S” and “D” symbols to indicate whether the linked text was searchable or could be downloaded. Regarding issues of access, users want access to full-text resources anytime and anywhere. When asked what they like most about e-books at their institution, 10 participants mentioned the high amount of titles available. Yet, when asked what they like least, problems with access were commonly cited. The most frequently reported issue was functionality. Users expressed the desire to see the functionality of e-resources indicated in links. This suggestion is related to the idea that users want a consistent interface, since many participants preferred a particular interface because of their familiarity with it. Familiarity with publishers was another instance of the general preference for familiar sources (familiarity with the source affects user decisions), since users tend to choose links based on their knowledge of whether that source offers the functionality or access they require. Finally, it became clear that wording of links has a high impact on user experience. Some labels were perceived as clear to most participants; for example, 10 participants said that “Electronic version via” was clear to them. However, many participants reported that the wording of other labels was either misleading or confusing. As an example of a misleading label, participants who clicked on “Ebrary Academic Complete” explained that they believed “complete” meant they had access to the full text, which was not always the case. Participants expressed confusion when faced with two links with exactly the same wording.

Overall, the findings revealed a relatively high acceptance of e-books not only for leisure but also specifically for academic purposes, and helped to identify specific usability issues with the partner institution’s interface for accessing e-books. However, the results did not provide insight into why some academic library users make use of e-books only occasionally or never and whether user expectations for e-books are influenced by their experiences with printed media, the nature and demands of academic tasks, or other factors. These research questions, scarcely addressed in the literature, will be investigated in the next phase of the study.
Conclusion and Future Work

This pilot reveals that a large percentage of students and faculty who participated in the study use e-books not only for leisure but also for academic purposes, and that they access them through their academic library. While most of the participants were generally pleased with their interactions with e-books, they reported having less satisfying experiences with the use of academic e-books accessed through the partner institution’s library compared to those acquired either through other libraries or personal purchases. As this study progresses, we will investigate the research routines of academic e-book users to identify factors affecting their preferences and patterns of use in context (e.g., project deadlines, academic year milestones, etc.).

While these results cannot be generalized due to the small size and the homogeneous nature of the sample, this pilot offers preliminary insights into the use of e-books in an academic environment and the expectations of its users. It also lays the foundation for future studies to broaden and deepen our understanding of the changing practices of accessing and using e-books.

References


Abstract

Building online communities to support small, audience-bounded offline social structures such as neighborhoods or organizations can be challenging. Due to the small size of their potential audience, the contribution volume is likely to be insufficient to maintain a sustainable community-driven system. In our research, we are interested in studying how different characteristics of the offline structure of these communities can influence their online behavior. Particularly, we analyzed participation of researchers in a social system for conferences. Our analysis shows that centrality in the academic social structure is a significant predictor of the likelihood to accept an invitation to participate in the system. These results suggest that an understanding of the users’ offline context can increase the effectiveness of user engagement strategies in an online context.

Keywords: audience-bounded online communities, offline structure, participation

Introduction

Online communities have been widely known for enabling communication among large numbers of people distributed around the world. However, online communities can also greatly benefit smaller existing offline groups. Providing virtual spaces for communication to existing offline groups can increase community awareness and engagement (Schuler, 1994; Carroll, 2012). These kinds of online communities aim to reach much smaller cohorts of people who share an offline context and have opportunities to interact outside the online community. For example, there are private online social networks serving neighborhoods (e.g. Nextdoor.com); online communities to enhance the experiences of the event attendees (Farnham, Brown, & Schwartz, 2009; Farzan & Brusilovsky, 2008); educational online communities to support students; and organization-specific social network sites to share knowledge and to increase networking among employees (DiMicco et al., 2008).

One of the main challenges for online communities is to maintain a sufficient level of contribution. Even with a potentially global reach, online communities often fail to maintain a critical mass (Cummings, Butler, & Kraut, 2002; Raban, Moldovan, & Jones, 2010). Building self-sustainable social online systems for already existing offline communities is even more challenging because of the much smaller population of potential participants (Resnick, Janney, Buis, & Richardson, 2010). Therefore, special efforts focusing on user engagement are required.

The shared offline context has the potential to affect members' motivation to participate in the online community. For instance, it can influence the perceived usefulness of the online community. The fact that the potential users of an audience-bounded online community are likely to know each other or share some offline context adds to the challenge. The perceived benefits of using the systems need to surpass not only the costs of participating as in every online community (Butler, 2001), but it also needs to provide benefits that cannot be easily achieved through the offline interactions. Our research agenda aims to understand how

Acknowledgements: Claudia Lopez's PhD studies are funded by CONICYT, Chile (The National Commission for Science Research and Technology) and a Fulbright Scholarship.


Copyright is held by the authors.
offline social properties of potential users are associated with the patterns of participation in audience-bounded online communities. To achieve this goal, we are collecting quantitative and qualitative data about the usage of online communities for conferences, graduate schools and neighborhood-oriented online communities.

In this paper, we analyze how offline centrality in an academic social network influences the likelihood of participation in an online community for academic conferences. Our results show more central members in the offline social structure are significantly more likely than peripheral members to respond to invitations to participate in the online community. Our findings reveals that the offline context of potential users can be used to design a more efficient user engagement strategies.

**Research Questions**

Economic modeling provides a framework for understanding peoples’ motivation in making decisions to take an action, including participation in online communities (Harper, Li, Chen, & Konstan, 2005). Often people compare the cost of performing an action with its benefits in order to decide to take the action. To participate in an online community, the perceived benefits of using the systems need to be more significant than the costs of participating (Butler, 2001).

A potential benefit of online communities is easier access to information (Schwabe & Prestipino, 2005; Cotten & Gupta, 2004). Online communities can provide more opportunities for individuals in terms of exposure to those who hold relevant information. Centrality in a social network has been found to be associated with access to more information (Brass, 1984). In this sense, we can argue that more central members of the offline community can get most of the information they need through their numerous offline connections while more peripheral members have less access to information because of fewer connections in the offline world. To compensate for this lack of information access, peripheral members can take advantage of online communities to acquire the information they cannot access offline. Thus, the perceived benefits of a social system can be higher for more peripheral members of the community. This leads us to our first hypothesis:

**H1a**: Peripheral offline community members are more likely to decide to participate in an audience-bounded online community than more central community members in order to compensate for their lack of connections offline.

On the other hand, research in organizational behavior suggests that more central members have a stronger sense of attachment to the organization (Burt, 2001). Their network position within the organization facilitates their understanding of what it means to be part of the organization and strengthens their sense of identification with the organization (Burt, 2001). In an online site for a community of practice, centrality in the online social structure was found to be correlated to higher contribution rates and more helpful contributions (Wasko & Faraj, 2005). In an offline context, people who define themselves as more knowledgeable and more invested in their local neighborhood wrote more reviews on social media sites than those who did not identify themselves as local experts (Antin, Sa, & Churchill, 2012). As a result, even though the personal benefit of contribution to the online community can be smaller for a central member, because of their stronger sense of attachment to the community, the perception of community benefit can be higher for them. This line of argument leads us to a competing hypothesis that central people are more likely to contribute because they have stronger attachment to the community and they care more about the community. Thus, an alternative hypothesis is:

**H1b**: Central offline community members are more likely to decide to participate in an audience-bounded online community than peripheral community members because of their stronger attachment to the community.

**Data Collection**

To test our hypotheses, we analyzed participation patterns of users in an online community for academic conferences. More specifically, we conducted an analysis of users’ behavior online in the Conference Navigator (CN) system during two conferences: the User Modeling, Adaptation and Personalization Conference (UMAP) 2011 and 2012.

Conference Navigator (Farzan & Brusilovsky, 2008) is an online community designed to help conference attendees make decisions about which talks to attend. CN users can schedule talks they want to
attend at the conference, and they can tag their own papers with relevant keywords to increase the visibility of their papers within the system. Based on user-generated content, CN guides conference attendees by augmenting the conference schedule with meta-information about talks such as popularity and by providing personalized recommendations of relevant talks to each user. Additionally, the users can connect to other researchers in the field through the system. Logging into the system is necessary to schedule talks, receive recommendations and connect with other researchers.

The system was initially introduced to the authors solely. The conference chair invited all the papers’ authors to preview the system before the conference. Two invitation emails were sent to each author. The first invitation letter invited all authors to schedule their five most interesting papers in the conference. The authors were told that this information will help the conference chairs to plan the schedule of the conference by identifying clusters of relevant papers and trying to schedule them for a single session during the conference. This first message was sent ten weeks before the conference. Six weeks later, after the conference schedule was released, the conference chair sent another email to the same list of authors. The authors were invited to check their papers’ information and inform the system administrator if they found any error in the paper and presentation details. We collected the data of the authors’ behavior in the system during the interval of time from the first email until a week before the conference, when other conference attendees were invited to use Conference Navigator.

To model the offline social structure among the conference attendees, a network of co-authorship was created by using the data about the papers published in the UMAP Conference series from 2001 through 2011. Co-authorship relationships are defined when a pair of authors has written a paper together for UMAP series of conferences. These relationships are weighted based on the number of papers two authors have written together, number of papers by each of the authors, and number of co-authors they had in common. We employed the method described by (Liu, Bollen, Nelson, & Sompel, 2005) to calculate these weights. We used Gephi tool\(^2\) to calculate network measures including clustering coefficient and degree centrality. The network of authors includes 491 authors. The average degree centrality is 4.74, the average clustering coefficient is 8.81, and the average path length is 2.46.

Data Analysis

We analyzed how the UMAP authors’ network properties were associated with their decision to participate in Conference Navigator. We used logistic regression to model these relationships. The dependent variable is a binary variable: “1” for the authors who participated in the CN system at least once and “0” for those who never logged into the system. The independent variable is the centrality of the author in the academic network.

It is possible that first authors perceive the system to be more beneficial as they may view the system as a medium to propagate information about their paper. This is especially important if our first hypothesis is true and personal gain is the main motivational factor for participation in the system. To account for that, we controlled for the fact that the authors were a first or a secondary author of a paper in the conference.

Beyond network centrality, degree of connection in users’ neighborhood can influence their perceived benefits of the system. If community benefit is what most motivates users, users whose neighborhoods are less connected may perceive that the system can help to increase neighbors’ awareness about each other and help them to connect to each other. To account for that, we also controlled for the level of connection among the author’s neighborhood by including the clustering coefficient.

Therefore, the covariates included in the model are:

- **Author Role** indicates if the user is a first author or secondary author of a paper in the conference.
- **Weighted Degree** measures the users’ centrality in the academic network.
- **Clustering Coefficient** measures the connectedness the users’ neighborhood. A fully-connected neighborhood has a value of “1” and a completely disconnected neighborhood has a value of “0”.

Table 1 and Table 2 illustrate the descriptive statistics of our data. Among the 491 authors, 16% participated in the system (78 authors: 51 first authors and 27 secondary authors, see Table 1).

---

\(^2\) [http://gephi.org/](http://gephi.org/)
Table 1
Descriptive Statistics of Covariates

<table>
<thead>
<tr>
<th>Participation</th>
<th>Type of author</th>
<th>No (ratio)</th>
<th>Yes (ratio)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>107 (.68)</td>
<td>51 (.32)</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>306 (.92)</td>
<td>27 (.08)</td>
<td>333</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>413 (.84)</td>
<td>78 (.16)</td>
<td>491</td>
</tr>
</tbody>
</table>

As shown in Table 2, the mean weighted degree among first authors is 1.67 and among secondary authors is 1.41. The data suggests that first authors and secondary authors are not different in terms of centrality in the community. Regarding the clustering coefficient, the mean value among first authors is .72; for secondary authors, the mean clustering coefficient is .87.

Table 2
Authors’ Participation in Conference Navigator

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Kind of author</th>
<th>Mean</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted degree</td>
<td>First</td>
<td>1.67</td>
<td>1.32</td>
<td>.99</td>
<td>11.25</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>1.41</td>
<td>1.15</td>
<td>.99</td>
<td>7.79</td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>First</td>
<td>.72</td>
<td>.41</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>.87</td>
<td>.29</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3 describes the results of the logistic regression. Members with higher centrality value in the offline network are significantly more likely to respond to our invitation to participate in CN after controlling for the other covariates ($p < .0001$). One unit increase in the weighted degree increases the chance of participation 1.4 times ($e^{.348}=1.4$). This result supports our hypothesis H1b that community benefit for central members is stronger motivating factor than personal benefit for peripheral members.

Keeping the network features constant, the first authors are 4.9 times ($e^{1.587}=4.9$) more likely to participate than secondary authors ($p < .0001$). This follows the general idea of cost-benefit analysis of participation that, independent of network centrality, the benefit of participation can be higher for first authors. They can use online networks as a medium to learn more about a community they care about and to promote their position in that community.

Members whose neighborhood is less connected are more likely to participate. However, this effect is only marginally significant.

Table 3
Regression Results - Authorship- Activity

<table>
<thead>
<tr>
<th>B</th>
<th>S.E.</th>
<th>z</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Author</td>
<td>1.587</td>
<td>.275</td>
<td>5.76</td>
</tr>
<tr>
<td>Weighted Degree</td>
<td>.348</td>
<td>.091</td>
<td>3.60</td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>- .595</td>
<td>.355</td>
<td>-1.67</td>
</tr>
<tr>
<td>Constant</td>
<td>-4.099</td>
<td>.603</td>
<td>-6.80</td>
</tr>
</tbody>
</table>

n=491, non-zero=78, $p <= .0001$

Overall and individual fit analyses were conducted to assess the robustness of these results. The Hosmer-Lemeshow goodness-of-fit test shows that there is no evidence of lack of fit of the model ($p = 0.43$). Some potential outliers were identified using the standardized Pearson residuals scores, Pregibon leverage and delta-beta statistics analyses. After removing these potential outliers, the interval confidence of the effect of first author fell below the positive range (lower than 1). However, centrality remained a significant predictor of the likelihood of participation in the online community. This suggests that the model is robust to outliers.
We observed the same result in a logistic regression to predict the likelihood to contribute (i.e. schedule a talk, tag a talk, or connect with others). More central members and first authors were more likely to contribute. We also analyzed the relationship between network centrality and contribution or participation levels in the system for those who logged in at least once. We conducted a robust Poisson regression with the number of contributions and number of actions as outcome measures with the same predictors. None of the models was significant, which may be due to the fact that a small percentage of users logged into the system.

Discussion and Limitations

This paper reports a study that aims to understand the relationship between centrality in an existing offline community and the likelihood of participating in a related online community. Our results support the hypothesis that higher centrality in the offline community significantly increases the odds to decide to participate in the online space for the community. This finding proposes that designers deploying an audience-based online community would benefit from understanding the social structure of the targeted offline community. Identifying central people and strategically engaging them earlier in the online conversation could be an effective user engagement strategy. Engaging more central people early can help to create enough initial content that can be later used to engage more peripheral members. Members that are less central to the community may need more reasons to join the system. Having valuable content contributed by central members and an active stream of activity can make the system more valuable for them.

We are interested in understanding why more central community members are more likely to join the online system. We argue that central community members are more committed to the community; therefore, they are more likely to embrace a new initiative that can provide benefits to the community as a whole. It is also possible that more central members are more comfortable publicly contributing to the system because they know the audience better. Although these are plausible arguments, further investigation is necessary to confirm these speculations. Our future work will focus on collecting qualitative data to examine what motivates offline central members in the UMAP community to participate in an online community.

Additionally, we will attempt to replicate the study within larger conferences. The current study focuses on a small-size community featuring a highly connected offline structure. Compared with well-known average of six-degrees of separation in random communities, the UMAP community has an average path length of 2.46; i.e. this community is more strongly connected than a random community. Therefore, our results may not be directly generalizable to other social systems that aim to support bigger and more loosely-connected social networks. Therefore, we believe that replication of this study in larger conferences or other offline communities are necessary to make our results more robust.

Despite the limitations, we believe that our research provides a gateway to enhancing users’ engagement strategies in audience-based online communities.

References


MASHing Metadata: Legacy Issues in OAI Harvesting From Three Digital Libraries

Michael Khoo
The iSchool
Drexel University
khoo@drexel.edu

Doug Tudhope
Faculty of Advanced Technology
The University of Glamorgan
dstudhope@glam.ac.uk

Ceri Binding
Faculty of Advanced Technology
The University of Glamorgan
cbinding@glam.ac.uk

Hilary Jones
Mimas
University of Manchester
Hilary.Jones@manchester.ac.uk

Ivan Orrego
The iSchool
Drexel University
ifo23@drexel.edu

Jae-wook Ahn
The iSchool
Drexel University
ja626}@drexel.edu

Abstract

This Note reports on efforts to build a generalizable OAI-PMH workflow to retrieve metadata sets from unrelated digital libraries. This effort is part of a wider effort to build a database to aggregate metadata from different digital libraries, which can then be used as the basis for content analysis and data mining experiments with the metadata records. A pilot metadata harvest from three digital libraries using OAI-PMH encountered a number of issues, arising from idiosyncratic legacy characteristics of each of the three metadata sets. In the end, the harvests had to be manually tailored to each library. OAI-PMH proved to be a useful approach, but only after communication with each digital library had identified important characteristics of each metadata set, including many legacy characteristics, which had to be accounted for in the harvest.

Keywords: digital libraries, Dublin Core, legacy issues, metadata harvesting, OAI-PMH

Introduction

Interoperability is both a desirable and also an elusive goal for digital libraries (Paepcke et al., 1998; Gradmann, 2009). Individual digital libraries contain collections of high-quality resources, and combinations of resources from separate libraries can yield rich educational and research insights, but at the same time, it is not easy to search across different digital libraries. Federated search or browsing services are often simply not available, and where they do exist, they can, for various reasons, be of limited utility. While there is thus enormous potential present in the large numbers of digital libraries created so far, very often this potential is 'locked up' in what might be thought of as individual library silos. There are number of solutions to the 'un-siloing' of these digital libraries, such as mapping and crosswalking metadata in different libraries to a common format which can then be stored in and queried from a central repository. In reality, these solutions can be complex resource intensive endeavors, which can produce mixed results (Khoo & Hall, 2010; Lagoze et al., 2006). It is therefore worth exploring further strategies for digital library interoperability.

The work described in this Note is being carried out by three teams in the U.S.A. and the U.K., and is exploring methods for increasing discovery across unaffiliated digital libraries without the use of metadata crosswalks (Digging into Metadata, 2011: http://research.cis.drexel.edu/digging/). The three digital libraries in the study are the National Science Digital Library (U.S.A.: http://nsdl.org/) (also including the Digital Library for Earth Systems Education, DLESE: http://www.dlese.org/); the Internet Public Library (USA: http://www.ipl.org/) (also including the Librarians’ Internet Index (LII); and Intute (U.K.: http://www.intute.ac.uk/).

Acknowledgements: This research is funded by IMLS under grant # LG-00-12-0457-12, as part of the Digging Into Data Challenge. Khoo, M., Tudhope, D., Binding, C., Jones, H., Orrego, I., & Ahn, J-w. (2013). MASHing metadata: Legacy issues in OAI harvesting from three digital libraries. iConference 2013 Proceedings (pp. 497-501). doi:10.9776/13263

Copyright is held by the authors.
Harvesting Metadata Records for Content Analysis

A central premise of the work is that human-generated metadata records contain information which can be aggregated and subject to machine content analysis. In this study, a content analysis of metadata records collected from three digital libraries is being used to support the generation of Dewey Decimal Classification ‘tags,’ which will then be added back to each metadata record, in order to enhance search and browse functionality across the three libraries.

Not all fields in a metadata record are equally useful for content analysis, and so the initial work has focused on selecting and analyzing the title, description, and keyword and subject fields, in a small sample set of 50 metadata records obtained from these three digital libraries. An outline of the analysis is provided in Table 1, which shows a hypothetical metadata record (column 1), the fields selected for further analysis (the Title, Description, and various Subject fields) (column 2), and the ‘cleaned’ content of these fields which will then be used for the content analysis (column 3).

Table 1
Metadata cleaning of a Dublin Core for a hypothetical website (“astrophysics.org”)

<table>
<thead>
<tr>
<th>Metadata obtained via OAI-PMH</th>
<th>Selected fields</th>
<th>Cleaned fields</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="">dc:description</a>A review of space science and astrophysics&lt;/dc:description&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="">dc:format</a>&lt;/dc:format&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="">dc:type</a>&lt;/dc:type&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="">dc:publisher</a>&lt;/dc:publisher&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="">dc:subject</a>astrophysics&lt;/dc:subject&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="">dc:subject</a>space&lt;/dc:subject&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="">dc:subject</a>science&lt;/dc:subject&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;/metadata&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Much of the work so far has focused on the development of a prototype tool – ‘Metadata Aggregation, Storage, and Handling’ (MASH) – that will aggregate and process the metadata from each of the digital libraries, before handing off results to a second tool (Document Indexing and Semantic Tagging Interface for Libraries: DISTIL) for the generation of DDC tags (Khoo et al., 2012). An overview of the MASH tool, identifying the initial points of data ‘hand-off’ between the different partners, is provided in Figure 1.

Figure 1. Overall workflow for the MASH tool.

Initial work with MASH used OAI-PMH to collect a small number of Dublin Core metadata records from each of the three ‘test case’ digital libraries; selecting the title, description and subject fields from each record in each collection; removing XML markup; and applying several content analysis methods to
different configurations of the cleaned metadata, in order to identify best methods for extracting sets of key terms for use downstream in the project. While it was thought that some of aspects of this work would be more complex than others, in practice, even apparently simple stages have had unexpected dimensions, including the initial metadata harvesting.

Legacy Issues Affecting Metadata Harvesting

The original project proposal envisaged developing a prototype harvesting workflow with the three pilot digital libraries, with a successful proof-of-concept paving the way for the development of a more generalizable OAI-PMH workflow, allowing the project to scale up to additional digital libraries. The MASH team began the work by carrying out several manual harvests of small metadata sets from each of the three libraries. A number of issues were encountered in this initial harvesting, including non OAI- accessible metadata, variations in subject fields, and undocumented metadata aggregation and ‘normalization’ in union repositories. The resolution of these issues required discussions with each individual digital library before the metadata itself could be harvested in a form that was useful for the project.

The first example is the IPL. This was founded in 1995 as an online reference service, and then began developing digital collections (Janes, 1998). Beginning in 2008, the IPL merged collections with the Librarians’ Internet Index (LII), and the IPL and LII metadata was crosswalked to Dublin Core and added to a Fedora database (Khoo & Hall, 2010). Three of the fields of interest to the project – dc:title, dc:description, and dc:subject – were placed in the main DC datastream. However, two further potentially useful fields from the original metadata were placed in other datastreams. First, theipl:subject field, which included custom-formatted IPL browsing metadata inherited from the legacy SQL metadata, was placed in a separate IPL datastream (this IPL datastream also archived all the original IPL metadata that could not be mapped to the 15-element Dublin Core set). Second, the somewhat complex relationships between item-level resources and various types of collections in IPL were mapped to an RDF triple, and placed in a third datastream (it was decided at the time not to use the dc: relation element, as the IPL did not have a consistent definition of the relationship between a collection and an item). There was therefore metadata useful for MASH that was neither visible nor retrievable in a standard OAI query to the Dublin Core datastream. Understanding the location and nature of this additional metadata required some familiarity with IPL history, in order to configure the harvest.

The second example is Intute, which was developed by a grass-roots community dedicated to online educational resource discovery (Joyce, 2008; Williams, 2006). Much Intute metadata was collected by previous partners and consortiums. Partly as a consequence, each Intute resource has both a DC record and also additional subject classification metadata stored in separate SQL tables, which can be drawn on as needed. These SQL tables are partly a legacy of the prior projects that were subsequently migrated to Intute, which utilized specific subject catalogs to suit the needs of particular audiences, for instance by offering domain-specific keywords. Once again, the harvesting issue here was that there was potentially useful metadata which were neither visible nor retrievable via standard OAI queries, and which required communication with the Intute cataloging staff, this time by email and teleconference, to locate and understand.

The third example is the NSDL, a federated multi-disciplinary STEM library, with a central metadata repository at nsdl.org. The central repository integrates metadata from the individual domain- specific portals, or ‘Pathways’ (e.g. Zia, 2004; Bikson et al., 2011), with Pathways’ metadata being passed via OAI-PMH. (The metadata harvested by MASH from NSDL has therefore been through at least two OAI-PMH pipelines – from the Pathway to NSDL, and from NSDL to the current project.) As NSDL Pathways are independent entities, the same resource can be cataloged in different ways by different Pathways. In these cases, the record displayed to users at nsdl.org is a ‘normalized’ version of all records created by individual Pathways, in which all the subject terms from each Pathway record were preserved (with editing for redundancy) in the ‘normalized’ record. This resulted in multiple (often similar) subject terms being displayed in the record, which could skew the results of the content analysis in MASH. Email communication with the NSDL’s metadata staff was required to clarify this situation, and to adjust the harvesting process for the current project.
Discussion

In each of the three cases just described, the various historical contexts within which each library created metadata had resulted in a number of legacy metadata issues. These legacy issues often came to light only by accident, as happened for example when manually comparing an XML record obtained via OAI-PMH from a particular library, with the Web display of the same record at the library’s Web site, and noticing discrepancies between the two. A useful finding from the initial harvesting experiments was therefore that it was not possible to create a single generic OAI-PMH query to retrieve all the required metadata from all the libraries. Additional metadata had to be accessed via refined queries, or sent as .sql or .csv files. This required additional communication with the metadata owners in order to understand how the metadata had been created and structured in the first place, before any action could proceed. On the one hand, therefore, while each library had contributed its metadata for the pilot MASH analyses in good faith, it was not obvious to the metadata owners exactly what metadata the MASH team were interested in; on the other hand, though, the MASH team did not have a full picture of what metadata was available. The team therefore had to engage in boundary spanning (e.g. Brown & Duguid, 2001), and perspective making and perspective taking (Boland & Tenkasi, 1995), between the needs of the MASH tool, and the specific characteristics of the metadata in each library (Table 2). This analysis and communication was time consuming.

Table 2
Examples of metadata issues that required communication with individual digital libraries

<table>
<thead>
<tr>
<th>Digital library</th>
<th>Metadata field</th>
<th>Issues</th>
<th>Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPL</td>
<td>ipl:subject</td>
<td>Legacy browsing data from previous version of IPL - not available in main Fedora ‘DC’ datastream - archived in ‘IPL’ datastream’</td>
<td>Email queries, face-to-face clarification</td>
</tr>
<tr>
<td></td>
<td>isMemberOf</td>
<td>RDF triple describing item-collection relationship - not available in main Fedora ‘DC’ datastream - archived in ‘NNN’ datastream’</td>
<td></td>
</tr>
<tr>
<td>Intute</td>
<td>classification</td>
<td>Domain specific classification keywords stored in a separate SQL table</td>
<td>Email queries, teleconference</td>
</tr>
<tr>
<td>NSDL</td>
<td>dc:title</td>
<td>In cases where multiple NSDL Pathways cataloged the same resource, the contents of metadata fields have been aggregated and ‘normalized’ into a representative record; the underlying Pathway metadata is still available but requires specialized OAI-PMH queries.</td>
<td>Email queries</td>
</tr>
<tr>
<td></td>
<td>dc:description</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>dc:subject</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the end, all the required metadata was harvested from each of the three libraries, and successfully added to the MASH database. One of the original assumptions of the project proposal – that it would be possible to create a generalizable OAI-PMH workflow to support the scaling of the project to other digital libraries – was not however supported by the pilot study. In the case of MASH at least, there was always a need to communicate with a library before a harvest was implemented. Nevertheless, while the technical workflow was not generalizable, other generalizations can be made from the three cases, especially with regard to the need for detailed organizational communication to obtain full descriptions of metadata sets before harvesting is implemented.
Conclusion

This Note has described attempts to use OAI-PMH to retrieve and aggregate metadata from three digital libraries. While OAI-PMH worked as a harvesting technology, variations in the metadata formats of the individual libraries required manual analysis and resolution before a satisfactory OAI query format was reached. In addition, some metadata useful for the MASH project was not available even through OAI-PMH. The roots of these issues were traceable to a variety of legacy factors that had shaped how the metadata in each library had originally been created, formatted, and stored. These issues then had to be clarified through communication with each of the digital libraries concerned. The workflow was thus more time-consuming than originally anticipated, although it was ultimately successful. These findings have useful implications for other projects seeking to harvest metadata from disparate digital libraries.

References


Certain Characteristics of iSchools v. Other LIS Programs

Robert Wedgeworth
Rutgers, The State University of New Jersey
rwedge@illinois.edu

Abstract

This exploratory dissertation study compares 17 iSchools and 36 Other Schools that offer the ALA-accredited Masters degree program with respect to certain characteristics. The study compiles quantitative and qualitative data on 32 variables and subvariables drawn from the 2010 ALISE Statistical Report, web sites of the schools, and the Elsevier SCOPUS database. Statistical analysis of the data reveals significant differences between the iSchools and the Other Schools. The analysis and results have been completed, but the conclusions and implications are still under consideration.

Keywords: ischools, qualitative analysis, quantitative analysis, information science, library science

Background

Beginning in the late 1980’s the deans of several of the leading LIS programs began informal meetings to share ideas and to coordinate their efforts. (Larsen, 2008) The deans, led by Toni Carbo Bearman at the University of Pittsburgh, sought to distinguish themselves from the broader LIS field. By 2002, the group had grown to include the deans of LIS programs in more than ten institutions and began to formally identify the group as ‘information schools,” or “iSchools.” (Larsen, 2008). By 2010 the iSchools had held its fifth annual iConference and counted 31 institutions within its ranks, seven of which were outside of North America.2 Seven of the 24 iSchools in North America originated from the fields of computer science and engineering rather than LIS programs. The iSchools’ web site (www.ischools.org) lists the primary membership requirements as $1 million in research funds annually; an emphasis on research; and a commitment to the iField. The principal organizing concept of iSchools’ programs is the relationship of information, people, and technology.

According to the 2010 statistics of the Association for Library and Information Science Education (ALISE), the iSchools movement now includes the majority of the largest full-time equivalent (FTE) faculties in the library and information science (LIS) field. (ALISE, 2010) The movement also includes the majority of those LIS programs in research universities. As the most significant effort to redefine LIS programs by advancing the concept of the iField, the iSchools are a logical focus for study. The problem is how to discern the differences, if any, between those programs that identify themselves as iSchools and those that do not (Other Schools).

The Problem

The purpose of this dissertation study is to identify certain characteristics of LIS programs that offer the ALA-accredited Masters degree in order to determine if there are significance differences between the iSchools and the Other Schools with respect to the faculty, students, curriculum, or resources. This study is based on data representing the academic years 2005-06 to 2009-10.3

Since it is axiomatic that faculty, students and the prescribed curriculum are the principal influences on students, the hypotheses for this study focus on:

Copyright is held by the author.
The problem addressed here is what “is” the LIS field and how do the iSchools differ from those that do not identify themselves as iSchools—if, indeed, they do differ. In the absence of an appropriate model for this research, the approach taken is to explore characteristics of the three major contributors to the outcomes of an academic program—faculty, curriculum and students. Utilizing both quantitative and qualitative data from multiple sources and exercising inductive reasoning, the study expects to develop a better understanding of the programs.

Relevant literature

Two studies by Burnett and Bonnici address the evolution of the LIS field to the iField. (Burnett and Bonnici, 2006) (Bonnici et al., 2009) Using the framework of theories advanced by Abbott (2001) these studies address the status of LIS programs and whether iSchools are advancing toward developing an iField. Wiggins and Sawyer (2012) survey iSchool faculty to determine diversity of research backgrounds Dillon (2012) explores the evolution of iSchools, including their emphasis on research and their relationship to LIS programs.

Methodology

This exploratory study compiles and analyzes data representing certain characteristics of 53 schools or colleges that offer the ALA-accredited Masters degree program, including 17 iSchools as of February 2010. The study excludes programs for which English is not the primary language of instruction; LIS programs located outside of North America; programs less than 15 years old; and those iSchools that do not offer the ALA-accredited Masters degree.

The data compilation source representing the relative size of the faculties, students, and the amount of funding support is the 2010 ALISE Statistical report. The data compilation sources representing the curriculum of the respective programs are the school web sites as of February 2010. The data compilation source representing the research degrees held by over 400 individual faculty members with professorial level appointments is the school web sites supplemented by Google searches. Source of the research records of individual faculty members between 2005 and 2009 is the Elsevier SCOPUS database. Five categories of research degrees in the data in this study are Education, Library and Information Science, Arts and Humanities, Social Sciences and STEM (Science, Technology, Engineering and Mathematics).

Eight categories of courses offered by the programs are used for this study:

- History, Issues and Policies
- General Management—including financial management and personnel management.
- Methods and Techniques
- Youth Library services
- Library Services—including archives
- Information Services
- Information Organization
- Information Management

The school web sites are the source additional curriculum data for this study addressing the overall scope of curricular offerings, including other Masters degrees, undergraduate majors and minors, advanced certificate, and Ph.D. programs. The master file of data sets for this study comprises data on
32 variables and subvariables. The study uses t-tests for the analysis of most of the data. However, logistic regression analysis addresses the data on curricular offerings and the data on faculty degrees. Since the t-test assumes a normal distribution of the data, additional analyses exclude extreme data points. Logistic regression analysis does not assume a normal distribution of data but can be sensitive to extreme data points. Additional logistic regression tests also exclude extreme data points. In all cases the secondary tests confirm the original results. Tables 1 and 2 display selections from the raw data.

FTE Faculty= FTE faculty teaching in the ALA Masters program
FTE Student= Total enrollment in the ALA Masters program
% All Student= Percentage of ALA Masters of total school enrollment
Income Total= Total school income
Income External= Total income from external sources

### Table 1
**Selected iSchool Statistical Data**

<table>
<thead>
<tr>
<th>Name</th>
<th>FTE Faculty</th>
<th>FTE Student</th>
<th>% All Student</th>
<th>Income Total</th>
<th>Income External</th>
</tr>
</thead>
<tbody>
<tr>
<td>iUBC</td>
<td>18.33</td>
<td>154</td>
<td>68</td>
<td>1587333</td>
<td>0</td>
</tr>
<tr>
<td>iUCLA</td>
<td>13.94</td>
<td>135</td>
<td>81</td>
<td>3628865</td>
<td>1526473</td>
</tr>
<tr>
<td>iDrexel</td>
<td>49.08</td>
<td>481</td>
<td>54</td>
<td>16339132</td>
<td>4675049</td>
</tr>
<tr>
<td>iFordaSt</td>
<td>32</td>
<td>304</td>
<td>50</td>
<td>8332771</td>
<td>2243941</td>
</tr>
<tr>
<td>Illinois</td>
<td>37.5</td>
<td>414</td>
<td>86</td>
<td>10645840</td>
<td>4221774</td>
</tr>
<tr>
<td>iIndiLIS</td>
<td>50.5</td>
<td>498</td>
<td>94</td>
<td>9592458</td>
<td>676325</td>
</tr>
<tr>
<td>iKentcky</td>
<td>12.75</td>
<td>157</td>
<td>100</td>
<td>1564598</td>
<td>101000</td>
</tr>
<tr>
<td>iMarylan</td>
<td>36.6</td>
<td>290</td>
<td>73</td>
<td>4558588</td>
<td>2016123</td>
</tr>
<tr>
<td>iMichign</td>
<td>32.51</td>
<td>363</td>
<td>88</td>
<td>18493938</td>
<td>5469334</td>
</tr>
<tr>
<td>iUNC</td>
<td>41.5</td>
<td>296</td>
<td>83</td>
<td>9906061</td>
<td>5496848</td>
</tr>
<tr>
<td>iNorthTX</td>
<td>28.8</td>
<td>503</td>
<td>86</td>
<td>9527136</td>
<td>6552976</td>
</tr>
<tr>
<td>iPittsbr</td>
<td>37.5</td>
<td>261</td>
<td>47</td>
<td>15633902</td>
<td>6314340</td>
</tr>
<tr>
<td>iRutgers</td>
<td>34.75</td>
<td>313</td>
<td>55</td>
<td>5864358</td>
<td>1632158</td>
</tr>
<tr>
<td>iSyracuse</td>
<td>57</td>
<td>117</td>
<td>11</td>
<td>28856482</td>
<td>2945502</td>
</tr>
<tr>
<td>iUTAustn</td>
<td>24.8</td>
<td>250</td>
<td>46</td>
<td>5261757</td>
<td>1236655</td>
</tr>
<tr>
<td>iToronto</td>
<td>29</td>
<td>377</td>
<td>77</td>
<td>7731147</td>
<td>431787</td>
</tr>
<tr>
<td>iWshngtn</td>
<td>49.67</td>
<td>282</td>
<td>48</td>
<td>16617517</td>
<td>11985299</td>
</tr>
</tbody>
</table>

### Table 2
**Selected Other School Statistical Data**

<table>
<thead>
<tr>
<th>Name</th>
<th>FTE Faculty</th>
<th>FTE Student</th>
<th>% All Student</th>
<th>Income Total</th>
<th>Income External</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>20.5</td>
<td>202</td>
<td>87</td>
<td>2238667</td>
<td>379489</td>
</tr>
<tr>
<td>Alberta</td>
<td>9.2</td>
<td>85</td>
<td>91</td>
<td>1853646</td>
<td>359631</td>
</tr>
<tr>
<td>Arizona</td>
<td>13.25</td>
<td>191</td>
<td>98</td>
<td>3844142</td>
<td>2221614</td>
</tr>
<tr>
<td>Emporia</td>
<td>11.56</td>
<td>227</td>
<td>96</td>
<td>3065001</td>
<td>2021111</td>
</tr>
<tr>
<td>KentSt</td>
<td>23.75</td>
<td>470</td>
<td>97</td>
<td>3597936</td>
<td>23</td>
</tr>
<tr>
<td>LIU</td>
<td>19.26</td>
<td>284</td>
<td>90</td>
<td>2781138</td>
<td>90000</td>
</tr>
<tr>
<td>LSU</td>
<td>10.5</td>
<td>121</td>
<td>88</td>
<td>1027992</td>
<td>115150</td>
</tr>
<tr>
<td>McGill</td>
<td>12</td>
<td>161</td>
<td>73</td>
<td>1884189</td>
<td>451271</td>
</tr>
<tr>
<td>Missouri</td>
<td>12.75</td>
<td>207</td>
<td>87</td>
<td>1900247</td>
<td>499947</td>
</tr>
<tr>
<td>NCCU</td>
<td>15.48</td>
<td>236</td>
<td>75</td>
<td>1671553</td>
<td>130296</td>
</tr>
<tr>
<td>Name</td>
<td># Full-time Faculty</td>
<td># Articles</td>
<td># Journals</td>
<td>Journal Cocites</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>---------------------</td>
<td>------------</td>
<td>------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>iUBC</td>
<td>10</td>
<td>39</td>
<td>25</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>iUCLA</td>
<td>12</td>
<td>42</td>
<td>25</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>iDrexel</td>
<td>25</td>
<td>275</td>
<td>159</td>
<td>1942</td>
<td></td>
</tr>
<tr>
<td>iFlrdaSt</td>
<td>18</td>
<td>146</td>
<td>46</td>
<td>293</td>
<td></td>
</tr>
<tr>
<td>illinois</td>
<td>32</td>
<td>205</td>
<td>87</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>iIndiLIS</td>
<td>21</td>
<td>204</td>
<td>88</td>
<td>530</td>
<td></td>
</tr>
<tr>
<td>iKentucky</td>
<td>9</td>
<td>40</td>
<td>23</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>iMarylan</td>
<td>23</td>
<td>292</td>
<td>115</td>
<td>591</td>
<td></td>
</tr>
<tr>
<td>iMichighn</td>
<td>39</td>
<td>443</td>
<td>215</td>
<td>708</td>
<td></td>
</tr>
<tr>
<td>iUNC</td>
<td>22</td>
<td>263</td>
<td>84</td>
<td>938</td>
<td></td>
</tr>
<tr>
<td>iNorthTX</td>
<td>21</td>
<td>152</td>
<td>50</td>
<td>195</td>
<td></td>
</tr>
</tbody>
</table>
Table 4
Selected Other School Research Data

<table>
<thead>
<tr>
<th>Name</th>
<th># Full-time Faculty</th>
<th># Articles</th>
<th># Journals</th>
<th>Journal Cocites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>11</td>
<td>30</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>Alberta</td>
<td>8</td>
<td>50</td>
<td>26</td>
<td>16</td>
</tr>
<tr>
<td>Arizona</td>
<td>13</td>
<td>38</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>Emporia</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>KentSt</td>
<td>21</td>
<td>173</td>
<td>105</td>
<td>39</td>
</tr>
<tr>
<td>LIU</td>
<td>13</td>
<td>29</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>LSU</td>
<td>10</td>
<td>65</td>
<td>18</td>
<td>86</td>
</tr>
<tr>
<td>McGill</td>
<td>10</td>
<td>88</td>
<td>40</td>
<td>93</td>
</tr>
<tr>
<td>Missouri</td>
<td>11</td>
<td>60</td>
<td>37</td>
<td>6</td>
</tr>
<tr>
<td>NCCU</td>
<td>8</td>
<td>21</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>11</td>
<td>45</td>
<td>22</td>
<td>34</td>
</tr>
<tr>
<td>Pratt</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Rhodels</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>StJohns</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>SouthFL</td>
<td>10</td>
<td>60</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>SUNYBuff</td>
<td>9</td>
<td>29</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>Tennesse</td>
<td>12</td>
<td>231</td>
<td>81</td>
<td>328</td>
</tr>
<tr>
<td>Albany</td>
<td>10</td>
<td>31</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Catholic</td>
<td>9</td>
<td>26</td>
<td>15</td>
<td>32</td>
</tr>
<tr>
<td>Clarion</td>
<td>12</td>
<td>12</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Dalhousie</td>
<td>11</td>
<td>35</td>
<td>23</td>
<td>13</td>
</tr>
<tr>
<td>Dominican</td>
<td>14</td>
<td>51</td>
<td>9</td>
<td>57</td>
</tr>
<tr>
<td>Hawaii *</td>
<td>9</td>
<td>76</td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>Iowa</td>
<td>8</td>
<td>20</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>NC Greens</td>
<td>10</td>
<td>12</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>Queens</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>San Jose</td>
<td>13</td>
<td>17</td>
<td>11</td>
<td>0</td>
</tr>
</tbody>
</table>
Results

The preliminary t-test analyses of the ALISE statistical data reveal the following:

- The FTE faculties of the iSchools are significantly larger ($M = 34.48$) than those of the Other School ($M = 17.53$).
- The average FTE ALA masters enrollments of the iSchools ($M = 305.59$) does not differ significantly from the average FTE ALA Masters enrollments of the Other Schools ($M = 246.39$).
- As a percentage of the total school enrollment, the average ALA Masters student enrollment of the iSchools ($M = 67.47$) is significantly lower than that of the Other Schools ($M = 86.50$).
- The average total income of the iSchools ($M = $10,244,816.65) is significantly greater than that of the Other Schools ($M = $2,813,549.53).
- The average external income of the iSchools ($M = $3,383,857.88) is significantly greater than that of the Other Schools ($M = $2,813,549.53).
- iSchools produce more research ($M = 256.35$) than Other Schools ($M = 49.41$) for the period studied.
- iSchool research is represented in more journals ($M = 115.06$) than the research of Other Schools ($M = 23.14$) for the period studied.
- There is a significantly higher level of journal cocitation among iSchool faculties ($M = 489.65$) than Other School faculties ($M = 47.26$) for the period studied.

Logistic regression analyses reveal significant differences in the pattern of courses offered for the ALA Masters in the iSchools than in the Other schools. In addition, there is a significant difference in the types of research degrees held by iSchool faculty members compared to Other School faculty members. Information organization and Information management were the most significant courses in predicting iSchool membership, while the STEM degrees were most significant in predicting iSchool membership.

Conclusion

Preliminary results indicate that iSchools are different from the Other Schools in size; in the patterns of their curricula; in the diversity of faculty research degrees; in the amount of research produced; the number of journals in which iSchool research appears; and the amount of cocitation among iSchool faculty. Interestingly, although the proportion of ALA Masters students to the total enrollment of the school is lower among iSchools than Other Schools, the ALA Masters enrollment is the dominant cohort of students in almost all iSchools. The level of research productivity within iSchools clearly confirms a greater emphasis on research than in the Other Schools. The pattern of curricular offerings in the iSchools is different from the Other Schools, but these data do not offer any clues as to why, since the number of courses prescribed for the ALA Masters is comparable in the two groups of programs.

Additional research on curriculum, interdisciplinary v. multidisciplinary research, types of STEM degrees, and types of external funding are indicated.
References


“Sometimes I believe as many as six impossible things before breakfast”: Outsider Naïveté as an Asset in the Initiation of Massive-scale Book Provision Efforts

Elisabeth A. Jones
The Information School, University of Washington
University of Michigan Libraries
eaj6@uw.edu

Abstract

Based on preliminary analysis of archival and interview data collected for a larger research project, this paper explores the role of naïveté regarding existing library practice in initiating four massive-scale public book provision efforts: the Boston Public Library, the Carnegie library program, the Google Books Library Project, and the Open Content Alliance. Based on these case examples, it is suggested that such naïveté can be a useful facilitator in the initial phases of such projects – and can sometimes make the difference between their materializing and not materializing – but that in order to succeed, these efforts must eventually draw in others with substantial domain expertise and/or ingenuity.

Keywords: libraries, digital libraries, mass digitization, paradigm shifts, access to information

Introduction

In The Structure of Scientific Revolutions, Thomas Kuhn describes several revolutionary scientific discoveries – such as Galileo’s analyses of pendulum motion and Dalton’s contributions to atomic theory – that were made possible in part by the discoverers’ intellectual position outside of the dominant paradigm of the scientific domain they revolutionized (Kuhn, 1970, pp. 119-120, 132-133). In these cases, a certain level of disciplinary innocence was a virtue: because they did not “know” that what they were doing was “impossible,” they were able to change the world. And arguably, this sort of phenomenon is not limited to the world of scientific discovery. In the world of library and information science, there have also been instances in which naïve outsiders have pushed forward projects of unprecedented – and even supposedly “impossible” scale – especially in the area of access to books.

In this paper, I explore four cases (two historical, two current) whose object was to provide as many books as possible, to as many people as possible, to a large extent free of charge, in which the primary funder and/or champion stood at a remove from existing paradigms of library practice. These cases – the Boston Public Library, the Carnegie library program, the Google Books Library Project, and the Open Content Alliance – form the basis of a dissertation project comparing their motivations, internal self-definitions, and initial structural and procedural implementations. For the two physical libraries, data collection involved extensive archival investigation, both in physical archives and online; for the digitization initiatives, confidential interviews were conducted at both of the host institutions (Google and IA), as well as several partner libraries (including all five of the original Google partners, as well as the University of California, the University of Toronto, and Woods Hole Oceanographic Institution). Interviewees were selected based on a combination of a snowball sampling methodology, and an assessment of their proximity to decision-making processes.

1 A full accounting of the contours of this data is available upon request. At the most basic level, for the physical libraries, three physical archives (at the Boston Public Library, Columbia University’s Butler Library, and the Library of Congress) were consulted, along with many primary and secondary sources retrieved from online repositories (mainly those created by the other two cases!). For the digitization initiatives, confidential interviews were conducted at both of the host institutions (Google and IA), as well as several partner libraries (including all five of the original Google partners, as well as the University of California, the University of Toronto, and Woods Hole Oceanographic Institution). Interviewees were selected based on a combination of a snowball sampling methodology, and an assessment of their proximity to decision-making processes.


Copyright is held by the author.
Though analysis of this data is ongoing, the outsider status of primary project facilitators has emerged as a preliminary theme. In this paper, I sketch the role of naïveté regarding existing library practice at high levels in each case, ultimately suggesting two points: first, although some naïveté may or may not be necessary to the initiation of large-scale information access efforts, it can certainly be helpful; and second, though naïveté can ease the way toward initiating a massive book-access project, the ultimate success of such projects depends heavily on the empowered involvement of individuals with significant expertise and/or ingenuity in working through all of the many details.

**Case 1: Boston Public Library (ca. 1852-1858)**

The Boston Public Library (BPL) originated many of the norms and systems that we now think of as being characteristic of the public library – including circulating collections (Hillard et al., 1876), offering both scholarly and popular works for broad public use (Second annual report of the trustees of the city library, 1854), and providing a skilled librarian able to assist patrons with myriad information needs (Board of Aldermen, 1857, citing Edward Everett). These innovations were necessary in the early history of the BPL because, although it was not the first public library in the United States, it was the first to serve a major metropolitan area, and would soon become one of the largest libraries in the country – two factors which also contributed strongly to its great influence over the public library movement as a whole (Ditzion, 1947; Shera, 1949). There is thus an extent to which the BPL’s leadership had no established public library paradigm to draw upon; no hard-etched standard practices to follow or reject. Still, even for this period, the BPL’s most significant donor – Joshua Bates, a Massachusetts-born London financier – had considerably less of an idea how such an institution might function than most of the others involved. Bates’s willingness to contribute, despite his own self-acknowledged lack of library expertise, played a significant role in making the BPL into the influential institution it would become. Bates provided an initial endowment of $50,000 and a few years later, $50,000 more for books to fill the library’s first dedicated building (Bates, 1852; Trustees of the Public Library, 1865). These donations had a tremendous impact. As one historian of the BPL suggests, “[b]efore his intervention there were words, after it there was a library” (Whitehill, 1956, p. 66).

The reasons for Bates’s offer had a great deal more to do with desiring to provide a healthy, book-lined place for poor boys like he had been to spend their evenings than with any specific notion of what an enormous public city library might actually look like or how it might function (Bates, 1852). Bates simply wished a library to exist, and for it to start off on a grand scale (Trustees of the Public Library, 1865). Yet, to Bates’s credit, though his donations could have made him extremely powerful in swaying the strategic direction of BPL in its early years, he never claimed any such control, preferring to leave it to the library’s Board of Trustees and others whom he felt knew better than himself. Throughout the rest of Bates’s correspondence with the library, phrases like “I’m satisfied to be guided by what you and Mr. Ticknor think best” (Trustees of the Public Library, 1865, pp. letter from Bates to Edward Everett, November 12, 1852) and “I rely on you, Mr. Everett and Mr. Ticknor, to put the matter right” (Trusted of the Public Library, 1865) appear with frequency, as do expressions of confidence in the skill and knowledgeability of the library’s local leadership.

**Case 2: The Carnegie Library Program (ca. 1880-1919)**

By the time Andrew Carnegie began his expansive program of library philanthropy a few decades later, the public library had taken sufficient hold as an institutional form for at least some paradigms of library practice to have come into being – drawing in no small part on the example set by the BPL. Still, Carnegie himself appears to have had little knowledge of or interest in these paradigms, except to the extent that their development allowed him to take for granted that they existed. Indeed, Carnegie’s rationale for his library donations – more than $56 million, providing 1679 library buildings in the U.S., among 2509 total worldwide (Bobinski, 1969, p. 31) – strongly echoes that of Bates for his donations to the BPL. At every library-related speech Carnegie gave, he repeated the same personal anecdote as a rationale for his philanthropy: when he was growing up, a man named Colonel Anderson opened his personal book collection to Carnegie and other boys, and the use of this collection made such an impact on Carnegie that he would later grandly conclude that “there is no human arrangement so powerful for

---

2 Everett and Ticknor were the two most influential members of the BPL’s founding board of trustees.
good, there is no benefit that can be bestowed upon a community so great, as that which places within the reach of all the treasures of the world which are stored up in books” (Carnegie, 1889b). And as the Carnegie Library program proceeded, it seemed quite clear, as MacLeod suggests, that Carnegie viewed his donations simply (or perhaps simplistically) as “continuations of Colonel Anderson’s modest collection” (MacLeod, 1968, p. 20).

Indeed, although Carnegie wielded somewhat more control over his system of library donations than Bates did over the BPL, he seems to have had an equally undeveloped idea of what libraries actually were and what resources they would need in order to function. But then, given the way that Carnegie structured his donations – he would provide money for a building only when a community had committed to providing both land on which to build it and tax revenue with which to fund its collections and services (Bobinski, 1969) – and given his views on philanthropy as articulated elsewhere (that is, that it should encourage self-sufficiency and independent effort by giving “nothing for nothing”) (e.g., Carnegie, 1889a, 1889c, 1920), it seems clear that he believed that the communities could – and should – work out the details of running the libraries for themselves. And indeed, Carnegie libraries, once established, tended to sink or swim based on the extent of local support and the relative talents of local leadership (Johnson, 1916; Macleod, 1968).


Moving into more recent history, the Google Books Library Project provides perhaps the clearest example of the beneficial naivé explored here. In 2001, Google began approaching major research libraries – starting with those at founder Larry Page’s almae matres, the University of Michigan (UM) and Stanford University – about their desire to scan all the books in the world (or at least many millions of them), starting with all of both universities’ book-based holdings. Their overtures were successful, and those partnerships, along with less expansive agreements with Harvard, Oxford, and the New York Public Library, were announced in 2004, with a few dozen more to follow. In the pre-announcement years, especially at UM and Stanford, several meetings and phone calls were held between the universities’ librarians and high-ranking officials at Google. Three interview participants for this study described in strikingly similar terms how, in those interactions, the librarians would bring forward myriad concerns from prior experience with book digitization, and the Google representatives would essentially wave them off. As one recalled,

I’d come in and we knew some things about duplication, how many copies of Nature we had, how many copies between the branch libraries in the campus, what percentage of the collection was duplicated. So I start to talk about that. … And Larry Page went “yeah, we don’t care about that.” Because it would be harder to sort that out than just do ‘em. And we went [meekly] “OK!”

As experts within the existing paradigm of digitization, it was difficult for these library leaders to see beyond all the potential problems; standing outside of that paradigm, the Google teams (perhaps hubristically), did not share their hesitation. And although many of the librarians’ concerns have certainly proven valid (especially with regard to copyright and metadata (e.g., Duguid, 2007; Samuelson, 2010)), as one of those librarians noted, Google’s naivé in these areas – and some librarians’ willingness to play along with it – helped make the project seem less impossible:

these Google people were like these adolescents with these great ideas. And we were like the old folks, in a way. [But we decided,] “when in doubt, let’s do what they want to do. Let’s think like they think. Let’s not think the way we think, and have to have everything has to be tight and neat and you’ve got to make sure you’ve got the holdings record right.”

By setting aside their strong sense from within the paradigm of librarianship that such scale was impossible – and specifically impossibly complex, impossibly expensive, and impossibly risky – Google’s library partners were able to take a significant leap. And, it should be noted, just because the librarians were willing to go along with Google’s model does not mean that they lacked input: in fact, my interviews at both Google and partner libraries indicate that Google relied critically on the expertise of partner librarians, especially at UM, Stanford, and Harvard, to help them work through the procedural details of scanning tens of millions of books.

511
Case 4: The Open Content Alliance (ca. 2005-2010)

The Internet-Archive-led “Open Content Alliance” (OCA) arose as a challenger to Google’s project in 2005, championing a more open, transparent approach to scanning (Kahle, 2005; Leetaru, 2008). Though the Internet Archive (IA) had by then been scanning and storing all sorts of media for many years, the OCA represented a massive leap in scale for IA in the book arena. Under the banner of the OCA, the IA sought to facilitate library book scanning by establishing scanning “pods” in various locations and by storing and providing free public access to digital books through the IA website (Kahle, 2007). And it quickly picked up influential partners, including the University of California, the Boston Library Consortium, and the Biodiversity Heritage Library (Open Content Alliance).

The driving force behind the OCA was IA founder Brewster Kahle, who did have a foot inside the existing library paradigm, having been scanning for years already. Still, there is also a sense from interviews with OCA participants that Kahle, coming out of the software industry, still had an outsider perspective. In particular, he felt that libraries moved too slowly, and although he believed the OCA could make things go faster, he had no clear plan for how the organization would independently function. As one partner librarian suggested, “he was just ‘Come on! Let’s just go scan some stuff!’ and there wasn’t really much thought of contract or MOUs, or you know, the usual stuff that define people’s relationships.”

And, placing a more positive spin on Kahle’s approach, one study participant suggested that the IA was in a liberated position…because we’re not beholden to the normal baggage that is associated with library systems and library practice, and none of us are librarians. We’re software people.

You know? And that allows us to sort of approach the issue in a totally fresh way.

And indeed, the IA has made solid strides in book scanning; the IA Text Archive (http://archive.org/texts) now contains more than 3.5 million items, and the IA continues to scan (and otherwise accumulate) books at a fair clip. Still, it bears noting that the Open Content Alliance per se no longer exists. And interview accounts suggest that to a significant degree, it broke down because of Kahle’s unwillingness to share control. As one early participant recalled,

Me and a few others, really early on, had a discussion with Brewster about ‘look, you can run this as part of your Internet Archive, and it’s an extension of your ego, and that’s fine… or you can establish it as a kind of a trust, which has its own independent governance… and that’s fine too. But you can’t do both.’

Kahle and the Internet Archive’s library-outsider enthusiasm may have been sufficient to launch the OCA, but they proved incapable of independently keeping it aloft, at least in that form.

Conclusion

Though each of these cases include unique manifestations of library-outsider naïveté, they are nonetheless similar in that in each case, at least one influential individual believed in the possibility of something unprecedented – a major urban public library, a constellation of library buildings across the globe, the world’s premiere research libraries made digital within a decade, open collaboration on distributed book scanning – and used their considerable financial and/or technical resources to help usher that unprecedented thing into existence. For the BPL and the Carnegie libraries, it seems likely that the involvement of Bates and Carnegie merely accelerated a process that would have eventually come to pass without them – though perhaps not at equivalent scale. For the digitization initiatives, however, this outsider involvement was more transformative. As many of the librarians interviewed affirmed, before Google announced its Library Project, the library community dismissed the idea of scanning all the books in the world as a ridiculous pipe dream. Eight years later, the world is still far from having digitized all the books – but with more than 20 million already online, and multiple large-scale book-scanning efforts now underway, it doesn’t seem nearly as inconceivable a goal. Still, the extent to which each of these projects has succeeded – or not – has depended crucially on the involvement of less naïve individuals in working out the actual specifics – to name one for each of the first three cases, respectively, Charles Jewett (Whitehill, 1956), William Brett (Bobinski, 1969, p. 31), and John Wilkin.\(^3\) Being innocent of existing library “truths,” it seems, can be a significant aid in dreaming big in the book-access arena; however, beyond a certain point, more detail-oriented expertise and ingenuity is required to transform those ambitious dreams into a new reality.
References

Board of Aldermen. (1857). Communication from the mayor, transmitting a memorial from the trustees of the public library, suggesting certain changes in the library ordinance. Boston.
Second annual report of the trustees of the city library. (1854). Boston: City of Boston.

1 John Wilkin is the University of Michigan’s Associate University Librarian for Library Information Technology; he was rated highly in terms of influence by seven out of the eleven interviewees involved with Google Books.
Augmenting Optical Character Recognition (OCR) for Improved Digitization: Strategies to Access Scientific Data in Natural History Collections

Deborah L. Paul  
Florida State University  
idDigInfo, iDigBio  
dpaul@fsu.edu

P. Bryan Heidorn  
University of Arizona  
School of Information and Library Science  
heidorn@email.arizona.edu

Abstract

The Augmenting OCR Working Group (A-OCR WG) at Integrated Digitized Biocollections (iDigBio) seeks to improve community OCR strategies and algorithms for faster, better parsing of OCR output derived from valuable data on natural history collection specimen labels. This task is exceedingly difficult because museum labels are often annotated, and vary in content, form and font. Under the National Science Foundation's (NSF) Advancing Digitization of Biological Collections (ADBC) program, iDigBio is building a cyberinfrastructure to aggregate quality data from museum specimens housed in collections across the United States for use by researchers, educators, environmentalists and the public. Since March of 2012, the A-OCR WG formed from community consensus to begin its role in this endeavor, defining reachable goals including setting up a hackathon concurrent with iConference 2013. This paper reports on the definition of some key problems identified by the A-OCR WG since these science problems will drive research and cyberinfrastructure development.

Keywords: iDigBio, OCR, natural language, information analysis, machine language

Introduction

iDigBio is a NSF Project under the Advancing Digitization of Biological Collections (ADBC) program. We at iDigBio are building a cloud-based cyberinfrastructure to aggregate United States vouchered specimen data across biological and paleontological collections. Natural history collections labels contain vital primary data about the specimens including for example, an assigned scientific name, location and date of collection, name of the collector, a collector identifier for the specimen, a museum identifier, and sometimes description of the specimen and the environment from which it was collected (NIBA, 2010) (http://digbiocol.files.wordpress.com/2010/08/niba_brochure.pdf).

This information has a broad range of scientific uses such as source data for ecological niche or historic species distribution models. The scope of this 10-year project requires innovation to succeed as it is estimated there are well over 2 - 3 billion specimens in the world (OECD, 1999; Ariño, 2010). NSF and the broader community recognizes the digitization processes currently in use to capture this data need to be faster and more efficient to meet difficult challenges facing science and society (Chapman, 2005; Blagoderov, 2012).

Acknowledgements: iDigBio and the efforts of the A-OCR WG are supported by NSF Award EF-1115210. For more about the origins of iDigBio and the ADBC program see A Strategic Plan for Establishing a Network Integrated Biocollections Alliance (NIBA, 2010). Many kind thanks to all the iDigBio Augmenting OCR Working Group members for contributions and hard work to date. Working Group members listed here in alphabetical order: Robert Anglin, Jason Best, Renato Figueiredo, Edward Gilbert, Nathan Gnanasambandam, Stephen Gottschalk, Elspeth Haston, Bryan Heidorn, Daryl Lafferty, Peter Lang, Gil Nelson, Deborah Paul, Nahil Sobh, William Ulate, Kimberly Watson, Qianjin Zhang. To the reviewers and iSchools organizers, we want to thank you for your encouragement, your interest in our topics and especially for your considered, thoughtful input.


Copyright is held by the authors.
Looking For Insights

The iDigBio Augmenting OCR (A-OCR) Working Group, one of several at iDigBio, formed with initial member suggestions from the broader community present at the October 2011 Kickoff Summit for iDigBio. Our working group seeks participation, collaboration and collective knowledge from a wider audience to address key issues in OCR use and structuring and correction of OCR output obtained from museum specimen labels as outlined below. We believe input from and collaboration with Information Scientists, Computer Scientists and Data Analysts is essential if we are to succeed in using OCR effectively for speeding up digitization, ensuring better quality data and effectively disseminating the information where it is needed.

Natural history collections and in particular biological and paleontological collections are composed of samples from that natural world and are meant to represent the world in a number of ways. They are not just collections of objects such as fossils, plants, insects, fish, birds, mammals or microbes but also collections of rich information about those objects that may include such specific details as habitat, elevation, soil type, host species data, associated species present, and even species not present (i.e., absence data). While perhaps more common in the past, many collectors also continue to use a field notebook for recording exquisite details about each collecting event (Canfield, 2012).

Inaccessible data. One difficulty is that many of these specimens were collected long before there were computer systems to track any of this associated information about these objects. Without a digital record, only a few individuals may ever manage direct access to any of this hidden data by going to the source museum in person. It is estimated that less than 10% of United States museum specimens have any online accessible record (NIBA, 2010). The NSF ADBC program was designed to help address this problem.

The OCR process is just one element of a longer process. In a typical workflow, a team decides which part of a collection should be digitized and then pulls these specimens from their storage cabinets and carries them to an imaging center. Here workers image the specimen and any label information. Current state-of-the-art OCR is well-suited to some common museum label and data types but not so successful with other types (Figure 1) where handwriting, spotted and yellowing paper, uneven text and other non-text objects confuse OCR.

Figure 1. Museum Specimen Label not suitable for current OCR. Yale University Herbarium. Used with permission.

Parsing. Some types of specimens such as some insects may have only a serial number pinned under the specimen. With current technology, OCR will likely not be cost effective for these and project staff may simply type in the numbers or use voice recognition. But, additional information about the specimen is often found in field notebooks and grey literature that are potential targets for OCR. Herbarium specimens too, among other specimen types, frequently have a rich collection of information affixed to the mounting material for the specimen. It is logical too, that collections housing a greater percentage of more recently collected material will have a higher percentage of labels where OCR canto...
produce usable output. OCR can be applied directly to the image of the specimen or the labels can be cropped and passed to OCR. The label in figure 2 exemplifies the type of information that can be found on one of these labels and shows the kind of label for which OCR can be effectively applied. Next, using the HERBIS / LABELX system (Heidorn, 2008), the resulting parsed and formatted OCR output from OCR of this label is shown.

Figure 2. Label suitable for effective OCR. Yale University Herbarium. Used with permission.

Parsed formatted OCR output of label in figure 2 from HERBIS/LABELX system.

<?xml version="1.0" encoding="UTF-8"?>
<labeldata>
  <bt>Yale University Herbarium</bt>
  <bc>YU.010782</bc>
  <in>Herbarium of Yale University</in>
  <hdlc>Plants of Puerto Rico</hdlc>
  <cn>Scientific Name: Adiantum latifolium</cn>
  <fm>Family: Q. Polypodiaceae</fm>
  <in>Common Name: Adiantum latifolium</in>
  <lcl>Locality: Mahoe plot 1-3, Rio Abajo State Forest</lcl>
  <hb>Habitat:</hb>
  <ftl>Comments:</ftl>
  <col>Collector: Mark Ashton and J.S. Lowe</col>
  <cdl>Date: 17 July 1934</cdl>
</labeldata>

The above is just one example of the processes our group seeks to improve in our efforts to get specimen data into databases faster. Unfortunately, OCR software is often not utilized to its fullest potential or the OCR output may be sub-par because the specimen label images are not in a format
appropriate for current OCR technology. For example, while handwriting recognition is an active field of
research, the success rate on a generalized collection of labels with handwriting from a variety of
collectors is near 0. In recent work, Steinke et al. (2010) have been able to digitize a high percentage of
the handwriting of the famous scientist, Alexander Von Humboldt because the system is trained for just
this one collector; this success is an exception rather than the rule for labels with handwritten content
since there are many thousands of collectors. Note the development by Steinke et al. (2011) of better
algorithms for recognizing non-text, non-handwriting elements may be of some use to algorithmically
remove image data that does not contain text or print.

The situation with type-written labels is brighter. OCR is fairly successful on modern typefaces
and well-aligned documents. Museum labels, however, are not in a standardized font or layout. Loose
typewriter pinions seem to have been the norm for decades in museums meaning letters are not
organized in straight lines, ribbon quality resulted in incomplete characters and age has led to paper and
print fading. Consequently, the OCR quality on some collections of type-written labels can be marginal.
The situation is exasperated by the fact that scientific vocabulary is not included in standard OCR
dictionaries. All these issues contribute to the need for careful research and development in OCR.

The text strings that are the output of OCR need to be parsed into individual elements and placed
into standardized formats for ingestion into databases or the semantic web. Much work is needed on
parsing algorithms to facilitate getting OCR output mapped programmatically to current data standards for
fast, automated information extraction and conversion into machine readable format (Heidorn, 2008; Ruiz
et al., 2009; Wei, 2012). Some collections contain large numbers of similarly formatted labels so it is
possible to write regular expression parsers to format the information. Most collections, however, have
much more variable formats. These issues combined with inherent OCR errors require the application of
more flexible approaches such as supervised or unsupervised machine learning approaches.

**Broader Goals of the A-OCR WG.** Members of our working group put together a current
summary wish list of topics to work on, see: [http://tinyurl.com/OCRHackathonWishList](http://tinyurl.com/OCRHackathonWishList). The working group is
collaborating to put together materials and consensus knowledge to help the community get more from
their OCR strategies. No one approach will work for all labels because of the idiosyncratic nature of the
collections but one goal of the A-OCR WG is to identify the methods which work best under different
conditions. Topics we are gathering material on include:

- known effective practices for getting the most from any OCR software,
- known issues that hinder good (useful) OCR output,
- reporting findings after working with real image data and programmers to improve parsing of OCR
output,
- lists of OCR software currently being utilized by the natural history collections community with
contact information,
- training and evaluation procedures and data for comparing methods,
- compiling OCR resources such as natural history dictionaries with scientific names, collectors,
location, institutions and other information,
- compiling relevant OCR related research in the iDigBio bibliography resource, and
developing user-interfaces and workflows for human-in-the-loop participation in parsing.

In addition, we also seek to identify opportunities to find and leverage existing tools and technologies
that are successful in and out of the biology digitization domain and find opportunities to integrate these tools,
or to seek funding for tool development.

**Our outreach strategy.** The A-OCR WG held its first in-person meeting, October 1 - 2, 2012 at
the University of Florida, home of iDigBio. The working group and invited guests initiated plans for our first
OCR hackathon being held concurrently with this iConference2013: Scholarship in Action: Data -
Innovation - Wisdom. Our first hackathon ([http://tinyurl.com/aocrHack](http://tinyurl.com/aocrHack)) concentrates on parsing and user-
interfaces in an effort to establish a baseline of what is currently possible and find more partners outside
our usual borders to move forward. In addition to this paper, we put together an iConference2013
Workshop titled, "Help iDigBio Reveal Hidden Data: iDigBio Augmenting OCR Working Group Needs
You" to formally introduce the iSchools community to iDigBio, this working group, and the digitization
efforts and challenges in the natural history museum collection world. In conjunction, on Friday, the A-
OCR WG will present an iConference2013 Alternative Event to report back on the hackathon which is
Wednesday and Thursday, February 13 – 14, at the Botanical Research Institute of Texas (BRIT).

**Sustainability.** As a community service and in order to begin creating a sustainable effort, plans
are in place at iDigBio to set up a permanent virtual OCR sandbox available at any time to facilitate OCR
Engine experimentation, and machine language (ML) and Natural Language Processing (NLP) algorithm improvement for natural history collections. This effectively provides a virtual hackathon for anyone interested in trying out new algorithms on a standard set of images, for example.

**Conclusion**

We are confident that through these combined collaboration and outreach initiatives, new partnerships will evolve leading to improvements in OCR strategies that will positively impact the outcome of this national effort to create a new data resource for everyone. Current research supports this (Haston, 2012; Steinke, 2010) and so we are looking forward to the challenge.

**References**


Memory Cues, Recall Strategies, and Alzheimer’s Disease

Lynne C. Howarth
Faculty of Information
University of Toronto
lyne.howarth@utoronto.ca

Erica Hendry
Faculty of Information
University of Toronto
erica.hendry@utoronto.ca

Abstract

This paper reports on partial findings from research exploring (1) how individuals with mild Alzheimer’s disease (AD) use memory cues in the form of representations (tokens) to recall life stories, and (2) ways in which representations (tokens) influence the nature and content of the recall narrative. Further, it examines whether memory recall differs in response to personal, participant-chosen memory cues, as compared to those selected by someone other than the participant. Reliance on personal artifacts used during two of three unstructured interview sessions resulted in recollections that seemed more scripted in delivery and circumscribed in detail. Researcher-selected tokens, used exclusively during session 3, yielded more fully formed recollections, and additional stories. Early findings suggest that generic associations may be at least equal to, if not more effective than, unique, individuated artifacts to engendering creative self-expression and vivid personal recall for those experiencing the initial memory loss of AD.

Keywords: memory cues, surrogates, recall strategies, sense-making, Alzheimer’s Disease

Introduction and Background on Study

Alzheimer’s disease, the most common form of dementia, is a progressive, degenerative disease of the brain. Dementia consists of symptoms that may include loss of memory, judgment, and reasoning, and changes in mood, behaviour, and communication abilities (Alzheimer Society of Canada 2007). Intelligence, educational level, occupation attainment, and (cognitive) leisure activities have been associated with reduced risk of dementia and cognitive decline (Bain 2006). Within this context, information could be seen as an important resource to preserving cognitive reserve, defined as, “increased numbers of synapses, or from an increased ability of the brain to cope with physiological insults” (Bain 2006, 247). Yet information science research focused on aspects of Alzheimer’s disease (AD) is decidedly scarce.

Theoretical Framing

To address this gap, a study engaging individuals with early-stage AD was undertaken to explore concepts of representation and memory within the framework of information sense-making (Dervin & Nilan 1986; Dervin & Naumer 2009; Savolainen 2003; 2007) and recall strategies. The study’s qualitative methodology does not clinically measure a participant’s memory function, but rather explores how memory recall may be influenced by representations acting as surrogates for some aspect (people, place, thing, event, etc.) identified in a participant’s personal narrative.

Psychological models of autobiographical memory (Conway & Loveday 2010), self-narrative (Baumeister & Newman 1994), and narrative structure (Bruner 2004) offer lenses for interpreting study participant narratives. The literature of material culture (Chaudhury 2002; Csikszentmihalyi &

Acknowledgements: The authors thank this study participant who gave generously of time and goodwill. The research project as a whole is funded by the Social Science and Humanities Research Council of Canada (410-08-1540). Howarth, L.C. & Hendry, E. (2013). Memory cues, recall strategies, and Alzheimer’s disease. iConference 2013 Proceedings (pp. 519-523). doi:10.9776/13267
Copyright is held by the authors.
Rochberg-Halton 1981) provides context to “objects” – the personal artifacts and researcher-chosen memory cues or representative tokens. From performance studies, embodied performative aspects of memory and narrative (Basting 2003a; 2003b) shed important light on interpretations of the individual self.

**Study Methods**

The research is exploring (1) how individuals with early-stage AD use memory cues in the form of representations (tokens) to recall memories of life stories, and (2) ways in which representations (tokens) influence the nature and content of the individual’s recall narrative. The study involves three sessions of unstructured interviews held across a number of weeks. The first elicits a personal narrative, while, during the second, the participant is invited to speak to 5-7 tokens (plus one wildcard) related to the narrative as selected by the researchers. During the third session, the participant is again asked to talk about the tokens from session two as well as to consider 2-3 new tokens chosen by the researchers to represent aspects of the personal narratives from the first and second sessions.

**Pilot Study (2011)**

In an ASIS&T Interactive Showcase Poster session (Howarth & Hendry 2011), the researchers reported on outcomes from an initial pilot study. Interpretations of the data encouraged our thinking that memory cues or surrogate tokens were useful adjuncts for stimulating recall in most cases, and, in some, eliciting even richer narratives or stories. The interview sessions also offered a kind of “neutral space” in which to engage in a safe, nonjudgmental, and social retelling of personal memories. The pilot study was useful to honing the participant-object narrative approach instrumental to the methodology applied to the larger project.

**Current Study (2012)**

This paper reports on the specific case of one participant engaged in a subsequent phase of the multi-year study. Unlike other participants, who interacted exclusively with researcher-selected tokens – consistent with the participant-object narrative approach – this one individual insisted on attending both the first and second sessions with a set of documents that were then used as prompts in a kind of scripted life narrative. The initial narrative followed a sequence determined by the order of the documents. Nonetheless the session yielded a series of “vignettes” sufficiently rich for the researchers to determine surrogates as representative tokens or memory cues for points within the narrative. For example, the participant told stories during the first session of repairing computers, meeting Queen Elizabeth II, and exploring training as a pharmacist. Figure 1 illustrates three surrogates chosen by the researchers that might serve as memory cues or representative tokens associated with the narratives. These included a microcomputer processor, a pill vial, and a matchbox with a picture of the Queen and Prince Philip.

![Image of tokens](image.png)

**Figure 1**

*Three Sample Tokens: microcomputer processor, matchbox with image of the Queen and Prince Philip, and pill vial.*
The invocation of the participant’s personal artifacts as memory cues during the first two sessions, and the invitation to engage with representative tokens chosen by the researchers for use in sessions 2 and 3, prompted exploration of a question additional to those of the overall study. In what ways does memory recall differ in response to personal, participant-chosen memory cues, as compared with those selected by someone other than the participant?

Findings

As with the pilot study (Howarth & Hendry 2011), the research found that, in most cases, the presence of tokens led to either "high resonance" associations expressed by the participant, or "new" resonance such as elaboration or additional associations or stories. As expected, the wildcard token did not evoke any resonance, and was, in fact, summarily dismissed by the participant. Two other tokens—a pill vial, and microcomputer processor—were consistent in evoking minimal resonance. In contrast to story "fragments" or "scripts" expressed in sessions 1 and 2, marked increases were seen in the participant's integration of new details into more "complete" stories in session 3, including such elements as back-story and present context.

As noted previously, the participant was eager to have his own artifacts (employment reference letters, newspaper clippings, and family photographs) on hand during the first two sessions, and at one point retrieved two additional artifacts in response to the conversation. In contrast, only those tokens selected by the researchers were engaged with during the third session. During session 2 the participant would respond to researcher-chosen tokens with brief narrative fragments, often turning to his own personal artifacts to illustrate the story and using a more scripted structure, or even to change the subject entirely. During session 3, and in response to researcher-selected tokens, only, full recollections and additional stories appeared to be more fully formed and detailed.

Discussion and Implications

While the participant could simply have felt more comfortable with the researchers during a third visit, and while this subset of a larger study deals only with one individual, there are suggestions of possible differences between participant-chosen, and researcher-selected memory cues. Reliance on personal artifacts resulted in recollections that seemed more scripted in delivery and circumscribed in detail. As with various reminiscence tools and activities, including personal memory boxes (Hagens, Beaman & Bouchard Ryan 2003), technically mediated forms such as “multi-media biographies” created and viewed with family members (Damianakis et al. 2009), and the use of automated video capture for future reminiscing (Crete-Nishihata et al. 2012), the interjection of individuated and uniquely personal artifacts, objects, events may evoke more "accurate" though somewhat “fixed” recall. The stories the participant retold or added in response to researcher selected representative tokens in session 3, would seem to underscore reminiscence tools and activities that engender creative abilities and self-expression, including TimeSlips™ (Basting 2003a; Basting 2003b), viewing art (Rhoades 2009; The Museum of Modern Art n.d.), and “generic” memory boxes (Reading Borough Council 2012). On the other hand, there was no way to authenticate some of the more vivid narrative from session 3, given that the participant had no access to his personal documentation. Thus, while stories were richer in their detail, there was no way of determining their relationship to “real” life stories from the participant.

This expression of doubt may say more about the researchers’ reliance on the “objectivity” of evidence provided within the participant’s physical documents than it does about the veracity of the individual’s narratives. In contrast to the earlier pilot study (Howarth & Hendry 2011) in which the researchers had no means of verifying any aspects of personal stories, this particular case provided opportunity — for better or worse — for assessing and calibrating “fact” or relative “truthfulness.” This begs the question of a kind of hierarchy of evidence based on perception. Expressed otherwise, when does the “truth” of documentation take precedence over the authenticity of narrative?

An awareness of this evidential hierarchy should make the researchers more sensitive to judging prematurely the integrity and authenticity of personal narrative. When memory and recall are themselves called into question by a clinical label, such as Alzheimer’s disease, denying the credibility of an individual’s personal reality as expressed in his or her self narrative, seems at least presumptuous, if not
unsound from a methodological perspective. Ultimately, for those with early-stage AD, their caregivers, and clinicians, questioning veracity of memory recall may be less important than the activity of engaging in personal storytelling that is highly social, inclusive, and reinforcing of both “the self” and individual cognitive reserve.

**Conclusion**

The broader study, still in progress, is considering ways in which cultural heritage institutions (libraries, archives, museums, galleries etc.) may play a role in fostering and facilitating personal reminiscence. In the same way that the Museum of Modern Art engages individuals with AD in group discussions about artwork (The Museum of Modern Art n.d.), and the Reading Museum creates “memory boxes” of cultural artifacts available for loan (Reading Borough Council 2012), are there opportunities for assembling customized “information memory boxes” containing objects that may evoke and reinforce ones life stories? Might the participant-object narrative approach prove useful in preparing kits that are identity-affirming for those experiencing some loss of self through progressive cognitive decline?

The specific case addressed in this paper, raised a question as to how memory recall differs in response to personal, participant-chosen memory cues, as compared with those selected by someone other than the participant. Findings suggested that generic associations may be at least equal to, if not more effective than, unique individuated artifacts in evoking memory recall. Cultural institutions may be reassured that the resources and objects from their collections may be as useful and important as personal artifacts, to engendering creative self-expression and vivid recall in those experiencing the memory loss of early-stage AD.

**References**


Do Members Converge to Similar Reasoning Styles in Teamwork?  
A Study of Shared Rationales in Small Team Activities

Lu Xiao  
Human-Computer Interaction Lab The University of Western Ontario  
lxiao24@uwo.ca

Abstract

Seeking better understanding of the group impact on the members’ reasoning styles, we have been using Rhetorical Structure Theory to analyze rationale statements collected from group activities. Our early analysis showed that in the studied activity context the participants all focused on subject matter relations with three most common techniques: providing contextual information of the main point (circumstance relation), offering additional details of the main point (elaboration relation), and evaluation of the main point (evaluation relation). The results also suggest that group is an influencing factor on members’ reasoning styles in terms of the rhetoric of the rationales, and the use of circumstance and elaboration techniques.

Keywords: reasoning style, group activity, Rhetorical Structure Theory, shared rationales

Introduction

Teams exist for various purposes. We are grouped together not only because we have shared goals but also we are expected to have better outcome than if we conduct the activity individually: when we work together we complete tasks more effectively and/or efficiently; when we learn together we have better learning outcome; and when we play together we have more fun. However, not only are these things different. As we interact with the other team members, we influence and are influenced by each other. How are individuals affected by team dynamics?

Our approach to understand this question is to study whether and how members’ reasoning styles are affected by the team. As a first step, we have been using Rhetorical Structure Theory (RST) to analyze the shared rationale statements collected from four collaborative learning activities by three student teams. The teams conducted these activities in their own virtual group workspaces. All the activities required students to articulate and share their rationales with their group in the virtual workspace. We compared the RST relations of the three groups’ rationales to examine the similarities and differences of these rationales at the semantic level.

Related Work

The impact of team dynamics on individuals have been studied in research communities of various disciplines. One focus is on the development of shared understanding among members. The term “shared understanding” is used here a very loose fashion as it can be referred to different terms in different disciplines. In Psychology this can mean team mental model that are shared and organized understanding and mental representation of knowledge about key elements of the team’s relevant environment (Klimoski & Mohammed, 1994). These elements include the team’s tasks, equipment, working relationships, and situations (Cannon-Bowers, Salas & Converse, 1993; Duncan, Rouse, Johnston, Cannon-Bowers, Salas, & Burns, 1996; Rouse, Cannon-Bowers, & Salas., & Salas, 1992). In Human Factors research community researchers also examine the development of shared mental model by measuring the convergence of members’ mental models (Kennedy & McCombe, 2010). Specifically,
Kennedy and McCombe proposed a framework for measuring the convergence of the mental models through the analysis of the team’s communication content (Kennedy & McCombe, 2010).

The “shared understanding” in teamwork can be referred to another important concept in Human Factors literature – shared situational awareness. Situational awareness is about one’s perception of environmental elements with respect to what is happening and what will happen in the near future, and how that will affect one’s status. Shared situational awareness refers to “the degree to which team members possess the same SA on shared SA requirements” (Endsley & Jones, 1997, p. 47; 2001, p. 48). It is argued in the literature that members develop shared situational awareness over time, which is critical for the smoothness of teamwork (Endsley & Jones, 1997, 2001).

In Education literature “shared understanding” in team learning activities indicates shared cognitive representations (Roschelle, 1992). A key relevant term is knowledge convergence - the collaborative learners’ mutual influence on their knowledge. One important assumption underlying collaborative learning models is that people exchange knowledge and converge their knowledge representation through interactions (Barron, 2003; Ickes & Gonzalez, 1996). Jeong and Chi operationalized the notion of knowledge convergence and reported their study of quantitative assessment of the amount of knowledge convergence occurred during collaborative learning (Jeong & Chi, 2007). Prior studies in Education literature maintain that learners who converge in knowledge benefit more from the learning activities than those who do not (e.g., Fischer & Mandl, 2005).

In this research program, the interested aspect of “shared understanding” refers to the convergence of the team members’ reasoning styles. The research question is then: as members work in a group do their reasoning styles converge in their explicitly articulated rationales?

Research Design

As a first step towards answering this research question, we analyzed the shared rationales collected from four activities of three student teams that are conducted in virtual group workspaces. In the following subsections we describe the corpus, our coding schema -Rhetorical Structure Theory, and our analysis procedure.

The Corpus

The three student teams belonged to a junior undergraduate project management class at a major US university. In this class, student teams researched best practices for managing distributed teamwork in each of its five project phases: project initiation, project planning, project execution, project planning, and project closure. The instructor designed five major activities corresponding to these phases. In each major activity, student teams first spent two days on a challenges assessment activity that had to be conducted in a virtual group workspace, and then one week on the rest tasks that could be completed through the teams’ preferred communication and collaboration approaches (e.g., face-to-face teamwork, email communication, etc). After all of the project phase activities were completed, the teams produced a final report and presented their work in front of the class.

The rationale statements were collected from the groups’ challenges assessment activities. In a challenges assessment activity, every team member was required to propose challenges that a distributed team might face and to provide rationales that justified the challenges. An example rationale statement is “You may have team members in other countries that have different customs and holidays. For instance, in Taiwan, when they celebrate the Chinese new year, they don’t work. You won’t be able to schedule work for this time”. This rationale was written to justify this challenge – “Creating a culture based work schedule”. After everyone finished proposing the challenges, the team was required to evaluate the challenges and select three most and three least important challenges that a distributed team needs to address. Each member was required to provide his/her selection and the rationale that justifies it. Then a team was required to provide the selection as a team and provide the team’s rationales. All these tasks took place in the groups’ virtual workspace. In the virtual environment, there was specific space for the members to provide their rationales. There was also a group chat to support synchronous and asynchronous communication among the members.

As aforementioned, each major activity had a challenges assessment activity. Therefore, there were five challenges assessment activities in total and they were identical except that they were for different project phases. The first activity was not included in the analysis because in that activity the
students were just introduced to the virtual environment and shown what considered to be good or bad rationales in the activity. In the current analysis, the corpus only included the members’ rationales that justified their proposed challenges. Of these collected rationales, 136 rationale statements were from team 1, 138 from team 2, and 122 from team 3.

Rhetorical Structure Theory (RST) is a theory of text organization created in the 1980s (Mann & Thompson, 1988). The theory aims at describing a text by annotating the text with a structured and well defined way. It views text as a collection of units and there are a variety of relations among these units. In other words, the text is a coherent piece that has parts playing different roles with respect to the meaning of the text and having relations to each other. RST noted these parts as nuclear and satellites. Nuclear refers to the essence of the text, and satellites are supplemental or additional information related to the nuclear. The theory identified three categories of relations between the parts: subject matter relations, presentational relations, and multinuclear relations. The writer’s intention of using subject matter relations is to convey the relation; while his/her intention of using presentational relations is to convince the reader to lean towards the nucleus. Multinuclear relations refer to the situation when a text has relations that do not carry a definite selection of one nucleus. Originally, the theory specified 24 relations (Mann & Thompson, 1988). The list has increased to 32 later (Mann & Taboda, 2012). Our coding schema used 32 Relations.

**Coding Schema - Rhetorical Structure Theory**

In RST analysis, one needs to identify the parts in a text, specify the hierarchical structure (at semantic level) of the parts, and then code the relations between the adjacent segments. Figure 1 shows an example of how a rationale statement was analyzed with RST schema. In this example, there are four segments (e.g., “My first …very much important”), three relations, and three hierarchical levels.

![Figure 1 An example of RST analysis](image)

The same text statement may be split to different segments and certainly may be interpreted with different hierarchical structure depending on how the analyst understands the semantic structure of the whole statement. Addressing this subjectivity introduced by the analyst, we had five researchers read the rationale statements with one of them read and analyzed it four times. We first had a research assistant A coded the rationale statements twice. Then she explained RST to two other coders and the three of them coded training data separately and held meetings to discuss the results. The three coders then coded the rationale statements independently. The investigator and the fourth research assistant next compared the analysis results and identified all the segments for the rationale statements that were analyzed differently (this could be due to the different hierarchical structure, different segment, or different relations between segments). These segments were reviewed and revised by the research assistant A and another coder afterwards. The research assistant A then coded the rationale statements of the agreed segments.
Results and Discussion

Similarity among the Participants’ Reasoning Styles

Our analysis shows that the participants showed a certain level of similarity in reasoning styles. This is based on the results from three aspects. First, as shown in Table 1, all three groups mainly used subject matters relations in all the activities.

Table 1
Percentage of Different Types of RST Relations

<table>
<thead>
<tr>
<th>Team</th>
<th>Presentational RST Relations</th>
<th>Subject Matter RST Relations</th>
<th>Multi-nuclear RST Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.6%</td>
<td>87.7%</td>
<td>7.7%</td>
</tr>
<tr>
<td>2</td>
<td>3.7%</td>
<td>88.9%</td>
<td>7.4%</td>
</tr>
<tr>
<td>3</td>
<td>3%</td>
<td>85.6%</td>
<td>11.4%</td>
</tr>
</tbody>
</table>

Second, the three RST relations that occurred most frequently in a team are also common among the teams. They are all subject matter relations: circumstance, elaboration, and evaluation. This finding indicates that the participants mainly used three types of additional information to help convey the main point in the rationales: the specific contextual information that situates the main point (circumstance relation), the additional details of the main point (elaboration relation), and the evaluation of the main point that demonstrates its core value (evaluation relation). Table 2 shows the percentage of these relations in coded relations in each team.

Table 2
Three Most Occurred RST Relations and Their Percentages

<table>
<thead>
<tr>
<th>Team</th>
<th>RST Relations and their percentages in the coded RST Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Circumstance (20%) Evaluation (17%) Elaboration (14.4%)</td>
</tr>
<tr>
<td>2</td>
<td>Elaboration (21.8%) Evaluation (18%) Circumstance (14%)</td>
</tr>
<tr>
<td>3</td>
<td>Evaluation (16.4%) Elaboration (13.7%) Circumstance (13.5%)</td>
</tr>
</tbody>
</table>

Third, RST relations that no one used in a team were strongly overlapped among the teams. Table 3 shows the number of RST relations that were not used by the participants and the commonality among the teams.

Table 3
RST relations that did NOT occur in the Rationales

<table>
<thead>
<tr>
<th>Team</th>
<th>No. of RST relations not appeared in the analysis</th>
<th>RST relations that were not used in all three team</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>Evidence, Justify, Motivation,</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>Restatement, Unconditional, Disjunction,</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>Multinuclear Restatement (7 in total)</td>
</tr>
</tbody>
</table>

Interestingly, five of the seven relations that nobody used are those with the writer’s intention of persuading or convincing readers (i.e., Justify, Motivation, Evidence, Restatement, and Multinuclear Restatement). As the activity required the students to write down the rationales in shared rationale spaces instead of group chat or other places explicitly for communication purposes, the students’ process of articulating the rationales could mainly be an individual reflective thinking process, instead of one in a dialogue that often has persuasion or argumentation purpose to the interlocutor. Another possible reason that the persuasiveness of the rationale statements was observed is the fact that we separated the challenges from their rationale statements and only analyzed the rationales.
The Team Impact on Reasoning Styles

Table 4 shows the percentages of presentational, subject matter, and multinuclear RST relations in each activity’s rationale statements of the three teams. The data show that all the teams have gradually used presentational relations more over time. This increase of the presentational relations’ percentage in all teams indicates that as the participants became more and more used to sharing the rationales in the virtual workspace, their processes of articulating the rationales may change gradually from just reflecting on their own to communicating with the others.

Table 4
The Percentages of Three Types of RST Relations

<table>
<thead>
<tr>
<th>Team</th>
<th>Activity</th>
<th>Presentational RST Relations</th>
<th>Subject Matter RST Relations</th>
<th>Multi-nuclear RST Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3.5%</td>
<td>87.2%</td>
<td>9.3%</td>
</tr>
<tr>
<td>3</td>
<td>5.7%</td>
<td>89.7%</td>
<td>4.6%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.9%</td>
<td>90.7%</td>
<td>5.4%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.2%</td>
<td>84.1%</td>
<td>10.7%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.6%</td>
<td>91.6%</td>
<td>6.8%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.4%</td>
<td>92.2%</td>
<td>5.4%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.9%</td>
<td>87.6%</td>
<td>8.5%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.7%</td>
<td>85.8%</td>
<td>8.5%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2.2%</td>
<td>84.9%</td>
<td>12.8%</td>
</tr>
<tr>
<td>3</td>
<td>2.3%</td>
<td>86.5%</td>
<td>11.2%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.4%</td>
<td>86.6%</td>
<td>11.0%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4.5%</td>
<td>85.6%</td>
<td>9.8%</td>
<td></td>
</tr>
</tbody>
</table>

Within a team, we did not observe any trend in the use of subject matters and multi-nuclear relations from one activity to another. We thus assumed that the activity did not play a significant role in the use of these two kinds of relations. And we conducted one-way ANOVA analysis to examine the team effect on the use of these relations. The ANOVA result showed that there was statistically significant difference among the teams with respect to the use of multinuclear relations (p = .04), suggesting that the team members have influenced each other on the use of relations that do not have a central focus, i.e., multinuclear relations.

The ANOVA result also showed that there was no statistically significant difference among the teams with respect to the use of subject matter relations. This result suggests that the teams have put similar weight on the relations that are intended for getting the main point across from the writer to the readers (i.e., the use of subject matter relations), which is reasonable given that the groups engaged in the same activities and the participants had similar level of experiences and domain knowledge. However, this analysis considered all subject matters relations together as one item. Would there be differences among the teams regarding the use of specific subject matters relations? To explore the answer of this question, we conducted ANOVA analysis on the three most occurred subject matter relations: circumstances, elaboration, and evaluation. We assumed that the groups did not develop a trend in terms of using these relations over time as we did not observe such patterns (see table 5). The ANOVA results showed that team was an influencing factor for circumstance and elaboration relations (p-value is .035 and .019), but not for evaluation relation. This result suggests that being in a team has affected one in the using the contextual information and additional details of the main point, but not the evaluation of the main point.
Table 5
The Percentages of Three Types of Most Occurred Relations

<table>
<thead>
<tr>
<th>Team</th>
<th>Activity</th>
<th>Circumstance Relation</th>
<th>Elaboration Relation</th>
<th>Evaluation Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>16.3%</td>
<td>18%</td>
<td>14.5%</td>
</tr>
<tr>
<td>3</td>
<td>21.7%</td>
<td>10.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>20.5%</td>
<td>12.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>19.9%</td>
<td>16.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>16.8%</td>
<td>16.4%</td>
<td>20.8%</td>
</tr>
<tr>
<td>3</td>
<td>15.2%</td>
<td>24.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>11.9%</td>
<td>20.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>16.2%</td>
<td>16.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>17.3%</td>
<td>10.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>11.0%</td>
<td>12.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8.3%</td>
<td>13.6%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion and Future Work

Various studies have been conducted to understand how team setting affects the individuals. This study is interested in one aspect of such effects, that is, the team’s impact on the members’ reasoning styles. We collected rationale statements from four group activities of three student teams. These rationale statements were articulated by the team members in a written format in the teams’ virtually shared rationale space instead of the group chat that was provided in the workspace to support synchronous and asynchronous communication.

Our early analysis results observed similarity in the participants’ reasoning styles because there were commonly non-used and used relations among the teams besides the teams’ similar emphasis on subject matter relations. The results also suggest that team is an influencing factor on reasoning styles in terms of the use of multinuclear relations and facilitating the articulation of the rationale’s main point. In this analysis, we’ve only looked at the three most occurred relations. More analysis is under way to understand if and how the group factor affects the use of other relations.

References

Educational Dialogue and Technological Utterances: 
A Phenomenological Approach to Developing a Theory of Evocative Design Practice

Michael Marcinkowski
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA U.S.A.
mrm5586@ist.psu.edu

Abstract

Online and distance education presents many challenges for both educators and system designers. The nature of these challenges is given particular clarity when examining online education in respect to pedagogical approaches which emphasize the importance of dialogue as part of the educational process. In order to begin to understand the dialogic role that technology plays in online education, a phenomenologically-oriented conceptual framework is sketched out. Taking computer interactions to be structured largely as interpretive processes that occur within a wider cultural context, the role of online educational systems in instructor-student dialogue is presented as a form of technological utterance, able to both sustain, as well as instantiate, messages in a dialogue between students and instructors. Introducing questions for future research, the aim is to develop design theory for both the design of online education as well as for design in general.

Keywords: cultural informatics, education, human-computer interaction

Introduction

Like countless other areas, education has felt the impact of the possibilities offered by the Internet and online communication. In many ways, the growth of online education presents a paradigmatic shift in educational practices and strategy (Harasim, 2000).

Such a move from a classroom-oriented paradigm to one which focuses on computer-mediated forms of communication highlights the role that dialogue and exchange plays in education. Particularly, in looking at online education, there is a constant concern for maintaining the value of such a dialogue and examining how it can be achieved in an online space (Beldarrain, 2006; Shea, Pelz, Fredericksen, & Pickett, 2002; Volery & Lord, 2000).

While didactic modes of education have been considered to be key for student success for some time (Freire, 1997; Mitchell, 1999), the challenge of facilitating dialogue in online spaces presents new questions about the nature of such a dialogue, particularly regarding the ability of online tools to support the type of rich, cultural interaction that is necessary for successful dialogue (Young, 2008). Such questions permeate both the possibility for dialogue between instructors and students, as well as the dialogue that occurs between students themselves (Beldarrain, 2006; Shea et al., 2002; Volery & Lord, 2000).

Acknowledgement[s]: Thank you to Fred Fonseca for his assistance with portions of the conceptual development presented here.


Copyright is held by the author.
The question of dialogue in design has been shown to bear more broadly to the relationship between designers and users in any endeavor, not just education (Buur & Bagger, 1999; Ulich, Rauterberg, Moll, Greutmann, & Strohm, 1991; Wright & McCarthy, 2010). As such, this paper intends to be a first step in laying out a mode of design thinking that is characterized by its dialogic nature, whether explicitly connected to education or not. Building from already established philosophical and theoretical approaches to design, the aim of this paper is to frame future research into online education design and to develop questions which may motivate future design research. While it takes its initial inspiration from pedagogical approaches emphasizing dialogue, the framing developed may be applied more broadly to any mode of design, whether there is an explicit aim for the facilitation of dialogue or not.

In this paper we will address the role that online educational tools and the design of learning materials can play in the building of dialogic communication between students and educators. We will examine the ways in which online tools themselves must be conceived of as an utterance within a dialogue. Looking first at the implications of a phenomenological understanding of computing in general, we will examine how the process of designing systems for online education may come to serve as an utterance to be interpreted in dialogue. Such "technological utterances" are seen to serve both the facilitation of a dialogic process and to provide tacit, cultural meanings that are themselves part of a dialogue between instructors and students. Finally, the transformative role that a process of translating traditional educational lessons into online materials will be discussed and future directions for research are identified.

Technology in the World

In coming to understand the modes of dialogue that online educational practices offer, it is important to address the ways in which such dialogic processes are supported by a wider background of cultural knowledge. In particular, given its oftentimes broad goals (Mitchell, 1999), education comes to most resemble culturally-based modes of computing which emphasizes the importance of the surrounding contextual and cultural milieu of computer use.

As a discipline, human-computer interaction (HCI) has moved steadily from an insular and Cartesian understanding of the relationship between users and systems toward an ever-more expansive view of a cultural and embodied user (Carroll, 1995; Dourish, 2004a; Suchman, 1987). Inspired in part by a phenomenological position developed by Martin Heidegger (2010), there has been an increasing conception of computer interaction as being something that takes place within a wider world, in terms of both a system's necessary function (Dreyfus, 1979, 2007; Winograd & Flores, 1987) and a user's interaction with a computer system (Chalmers, 2004; Dourish, 2004b). In this way, computers and users are both conceived of as inhabiting a shared world, with this sense of co-habitation being that which engenders both the possibility of use, as well as the contextual function of systems. Diverging from a simple technological rationalism, human-computer interactions come to be seen as being engendered by wide social and contextual factors, with research taking on a distinctly phenomenological character (Winograd & Flores, 1987). Information technology becomes not just a conduit passing messages from one place to another, but instead becomes entangled with such messages itself, both in its own function and in our use.

In addressing both the wider conception of the user and the increasing variety of factors that need to be addressed in the practice of computer use, researchers have increasingly pointed to not only the ability, but also the necessity of designing systems such that they are able to be engaged with as objects to be interpreted rather than simple tools to be used (Dourish, 2004b; Sengers & Gaver, 2006). Application design comes to function as a culturally-embedded dialogue between designers and their users (Gaver, Boucher, Pennington, & Walker, 2004; McCarthy, Wright, & Cooke, 2004). For designers, and educational designers in particular (Young, 2008), it becomes necessary to be able to understand how systems can come to be known as such evocative things once they are set into the world at large. Understood phenomenologically, information technologies come to be seen not simply as a tool or a method of communication, but as something with its own being and status in the world. Online educational systems come to not only facilitate a dialogue between instructors and students, but also serve as messages themselves within a wider, already existent dialogue between learners and instructors. In functioning as a message, designed systems take on the form of a technological utterance which is able to support communication, while at the same time presenting a message of its own (however tacitly expressed). It is necessary to approach educational software not simply as a tool set that
is able to fit a specific set of requirements, but rather as something capable of being evocative, meaningful, and subject to interpretation.

Use, Education, and Dialogue

With computing coming to be understood as standing as one phenomena among others, embedded within a wider world of cultural intelligibility, it is necessary to look at the ways in which technologies can be understood to play a part in dialogue between designers and users, instructors and students.

While much previous theoretical work on the topic of dialogue has sought to make explicit the role of language in dialogue (Freire, 1997; Gadamer, 2004; Heidegger, 2010), as the interpretive possibilities of HCI have developed, it becomes necessary to look at the way in which technology itself becomes something to be interpreted as message, rather than just as a simple mediating tool between two signal-processing users. In computer-based interactions, the design of applications and systems comes to stand beside language as vehicle for the passing along of meaning, and is, in turn, subject to a cultural interpretation in the same way as language. Computing and the breadth of meanings that it takes on becomes seen to be subject to the possibilities of basic intelligibility as given over by the cultural background of such computing.

Applications and systems can be understood as technological utterances that, laden with meaning, need to be accounted for in a dialogic process between educators and students, and more generally, between designers and users. Online educational platforms and their associated learning materials are not simply a medium or channel through which instructors carry on linguistic dialogue with students, but are themselves part of instantiating that dialogue. As a simple example: the inclusion of a glossary of terms alongside other materials in an online course highlights to students the importance of a certain vocabulary in a course. In designing such educational systems, there is a need to understand that dialogue with students begins with the system itself, and not just the explicit linguistic messages that are passed through it. This may even go so far as to indicate that the mere fact of a course being offered in online form expresses to students something about the course.

Implications for Educational Design

Understanding these dual roles that online educational software can play, both as conduit and as message, there are two distinct challenges for instructional designers. First, as any system being used in the course of distance education is itself seen as figuring directly in the dialogue between students and instructors, questions regarding the way in which instructors translate their lessons into an online form are raised. Second, as both the design and subsequent use of a system engender a certain mode of interpretation, there is the question of how systems designed for online education may adequately support not only dialogue between instructors, but dialogue between students as well.

Translating from education to technology

The hermeneutic philosophy of Hans-Georg Gadamer (2004; p. 388-391) presents a framing of the question of translation that proves to be useful when confronting the question of translating from one context to another. In online education, there may be a translation from a real-world, classroom-based setting to one that is technological and online, or simply from the idea that an instructor has regarding a course to its online implementation. While the traditional aim of translation may be to reproduce the exact meaning of a text from its original language to another, such striving for exactness is never possible. For a translator, they are neither able to appreciate the full and exact historical and cultural context of a work and as such are themselves unable to understand the original meaning, nor are they necessarily able to appreciate the wider context of the new audience for whom the translation is being prepared. There is always an element of compromise.

For educators and designers, there is an analogous process as they move from either the pure idea of a lesson, or from an already established classroom-based lesson, to designing an online course. There is an inevitable shift in the meaning of any lesson as it passes through a process of translation to technical realization (Muller, 2004). Even in the case of simply conceiving of a lesson wholly within the space of online education faces this sort of challenge of translation. In attempting to formalize any type of
message into a technical system, there is the constant challenge of how to represent the intended meaning. While such a difficulty is faced in any form of pedagogy—online or not—the need to structure lessons into a formalized language suitable for online use adds further complication.

Opening Technological Utterances to Other Voices

While the process of translating learning objectives to formalized, technological systems is already fraught with challenges, additional complications are brought to bear as the situation of online learning is conceived of one that is best envisioned as a dialogue not only between learners and students, but also dialogue between students themselves (Beldarrain, 2006; Shea et al., 2002; Volery & Lord, 2000). With this consideration in mind, there is a need to not only translate lessons into formal, technological systems, but further, lessons need to be translated into systems which can then be appropriated, interpreted, and taken over by students in service of their own inter-student dialogues. That is, the initial utterance that is an online learning platform itself needs to be such that it can be re-made and re-factored into an utterance to be put forward by a student for other students.

Conclusion and Future Work

In understanding our interactions with technology in a phenomenological way, a fundamental tension in online education is revealed. Given strategies which are meant to engage students in dialogue with instructors and other students (Beldarrain, 2006; Shea et al., 2002; Volery & Lord, 2000), how is it possible to design online educational systems which are able to accommodate both instructor-driven dialogue and inter-student dialogue? The key, it seems, comes in understanding how the meaning of any technological system is something to be interpreted and is something that is reliant on a tacit background of cultural intelligibility. In this way, the various backgrounds of students and educators comes to function as a sort of fulcrum in any such exchange, in both the process of dialogue and the process of translation.

The development of this conceptual framework of design and dialogue points to three distinct challenges:

1. How is it possible through the design of online educational systems to bridge the gap of cultural intelligibility that separates instructors and a diverse and potentially unknown population of students?
2. How are lessons translated from an idea conceived by an instructor to a formalization of that idea in a technological system?
3. How can technological utterances initiated in the course of an instructor-student dialogue be re-figured and re-translated in order to function as part of an inter-student dialogue?

Even in the case of technologically mediated online and distance education, it is still possible to focus on process of dialogue and the co-creation of meaning between students and instructors. In many respects, while some channels of robust communication between students and instructors (such as face to face communication) may be in large part closed off in the course of distance learning, other new and different channels of dialogue are opened up. This paper has focused on the ways in which the technological tools and online distance learning platforms themselves need to be seen as taking part in the process of educational dialog as technological utterances, and how their creation and use can be productively conceptualized as a process of the translation of inner meaning toward material utterance and interpretation.

References


Comparative Analysis of User Searching in Domain-Specific and Domain-Independent Digital Libraries

Oksana L. Zavalina
College of Information
University of North Texas
Oksana.Zavalina@unt.edu

Elena V. Vassilieva
College of Information
University of North Texas
Elena.Vassilieva@unt.edu

Abstract

This note reports preliminary results of a comparative transaction log analysis study of user searching in two digital libraries in the United States — a large-scale digital library in the domain of US history and a domain-independent federal-level digital library that aggregates digital collections regardless of their subject scope. This investigation reveals similarities in user search behavior among the two types of digital libraries with regard to rate of collection-level search use and application of search limits, most often occurring search categories, etc. At the same time, notable differences are observed in the rate of fielded search and phrase search use, average search query length and frequencies, and distribution of some search categories. This study provides empirical data to support digital library developers’ decision making regarding audience-based information organization in large-scale digital libraries.

Keywords: digital libraries, domain-specific digital libraries, domain-independent digital libraries, information search, search queries, search categories, transaction log analysis

Introduction

The United States funding agencies have actively supported digitization of collections of information objects — documents, letters, photographs, learning objects, etc. — of high cultural, historical, and educational value. For example, the Institute of Museum and Library Services (IMLS) has awarded grants to fund creation of over 600 digital collections. The National Science Foundation has funded over 300 digital collections (Skog, McCourt, & Gorman, 2009).

Large-scale digital libraries provide centralized access to hundreds or even thousands of such digital collections and millions of information objects. A number of large-scale digital libraries of varying subject focus and scope exist. One of the largest national-level digital libraries — IMLS Digital Collections and Content (IMLSDCC http://imlsdcc.grainger.uiuc.edu) with 650 digital collections — is aimed at a general audience and covers a wide range of subject areas and disciplines. IMLSDCC is representative of the most widely occurring type of a digital library — a domain-independent digital library. Some other national-level large-scale digital libraries are domain-specific, i.e., are created for distinct audiences. For example, Opening History (http://imlsdcc.grainger.uiuc.edu/history) with over 1500 digital collections is developed for researchers and educators in the domain of United States history.

To make sure that large-scale digital libraries successfully meet information needs of their intended user communities — either broad as in case with IMLSDCC or specific as in case with Opening History — the design of their respective discovery and access systems should be informed by user tasks such as finding, identifying, selecting, and obtaining information (IFLA, 2008).

People engage in information search to satisfy their information needs. Searching is one of the two major types of interactions between users and discovery and access systems such as library catalogs, databases, search engines, or digital libraries (Wilson, 2000). Searching is expressed through search queries — sets of one or more symbols (e.g., words, phrases, etc.) used to instruct a discovery and access system to locate potentially relevant information.
Transaction log analysis — “the study of electronically recorded interactions” between discovery and access systems and “the persons who search for the information found in those systems” (Peters, 1993, p. 41) — is one of the methods actively used for unobtrusive observation of user interaction with information systems, in particular, user searching. For example, Jansen, Spink, and Pedersen (2004) compared search query length and Boolean usage rates in catalogs and search engines; Beitzel and colleagues (2007), Jansen et al. (2007), Koshman and colleagues (2006), and Spink et al. (2002) categorized web search queries into groups such as people, places, and things. However, studies of user searching have largely focused on web search engines or online library catalogs while ignoring large-scale digital libraries that have been created with the goal to support information needs of researchers and educators in various domains.

To improve user interaction with large-scale digital libraries and facilitate access to information, needs and information searching of their users should be taken into consideration in systems design and development. Different groups of users may interact with such libraries differently based on their varying information needs. This may require different policies regarding organization of information in domain-specific and domain-independent large-scale digital libraries. However, systematic investigation into the user searching in the large-scale digital libraries, let alone comparative analysis, is in its infancy. Several studies (e.g., Agosti et al., 2007; Khoo et al., 2008; Pan, 2003) have analyzed transaction logs of large-scale digital libraries but they focused mainly on selection of search options and quantitative characteristics such as search query frequency, length etc. Only two studies (Zavalina, 2007; 2011) undertook detailed content analysis of user search queries in a large-scale digital library. However, user search queries in the different types of digital libraries aimed at different user communities have not previously been compared.

The study, preliminary results of which are reported in this note, addresses this gap through comparative analysis of user searching in domain-independent and domain-specific digital library.

**Methods**

This study analyzed transaction log data from a representative domain-independent large-scale digital library (IMLSDCC) and a representative domain-specific digital library (Opening History, hereafter referred to as OH). At the time of analysis, both digital libraries were hosted on the same server by the University of Illinois at Urbana-Champaign1 and, with exception of a color scheme, had identical user interfaces, which, among other options, included a rarely found option for collection-level search — a search where the user can search for entire digital collections as opposed to individual items. This similarity of interfaces makes the two digital libraries uniquely positioned as perfect candidates for comparative analysis of user searching as any differences in searching cannot be attributed to the differences in search interface.

This study aimed to answer the following research question: How does user searching in digital libraries that have wide subject scope and serve general audience (i.e., domain-independent) compare to user searching in digital libraries with distinct subject foci such as US history (i.e., domain-specific)? In particular,

- in search query length and search query frequency?
- in the distribution of search options: advanced search, collection-level search, etc.?
- in the distribution of search categories, such as personal and geographic names, dates, etc.?

One complete year of search log data, collected by Google Analytics application, for each of the two target digital libraries was used in this analysis. Non-empty search queries were grouped with identical queries. This resulted in a total of 5,917 unique search queries: 2,715 in OH and 3,202 in IMLSDCC. This sample size allows for generalizations with 95% confidence level and 5% margin of error.

Both quantitative and qualitative characteristics of search queries were assessed. Query length was measured as the number of words in a query, query frequency — as the number of times a query appears in the sample (cf., Spink et al., 2001). Unique search queries were analyzed qualitatively. They were categorized into ten search categories, including seven derived from the Functional Requirements

---

1 On August 1, 2012, the interfaces of two digital libraries were merged into one, under the IMLSDCC name.
for Bibliographic Records (IFLA, 1998; 2008) (work, person, corporate body, concept, object, event, and place), one from Functional Requirements for Authority Data (IFLA, 2009) (family), and two from the study by Zavalina (2007): class of persons, and ethnic group. Polysemic user search queries and most phrase queries were assigned to multiple categories. Whenever possible, search queries formulated in foreign languages were translated and categorized into appropriate search categories. This approach allowed to meaningfully categorize more than 95% of the queries in the sample. Approximately 4.71% of unique search queries in two digital libraries (4.27% in the Opening History and 5.09% in the IMLSDCC) could not be categorized into any of the search categories. These unknown category searches were excluded from further analysis.

A detailed coding manual had been developed to support coding activity; the manual included definitions and examples for each of the coding categories, along with other guidelines. In the coding process, the two authors or this note worked independently of each other and applied the same coding instructions to the same subsets of the units of analysis. The first author coded all 2,715 unique search queries in the OH sample; half of them were also coded by the second author. The second author also coded all 3,202 of IMLSDCC unique search queries; half of them were coded by the first author of this note.

To establish the reliability of the coding measures, there is a need to assess the amount of agreement among the coders. An intercoder reliability agreement coefficient of .90 or greater is considered acceptable to all and one of .80 or higher is acceptable to most situations. In this research, a strong intercoder reliability — 99.40% or Cohen’s Kappa of .976 for the OH dataset and 98.92% or Cohen’s Kappa of .954 for the IMLSDCC dataset — was observed between the two coders.

**Findings and Discussion**

**Search Query Lengths and Frequencies**

In the OH, search queries had an average of 2.32 and a median of 2 words per query (Table 1). The average length of IMLSDCC search queries was lower (1.94 words per query) while the median was the same. The median search query frequency was identical (1) in the two digital libraries. However, the mean was higher in OH (1.91) than in the IMLSDCC (1.60).

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>median</th>
</tr>
</thead>
<tbody>
<tr>
<td>OH search queries:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>2.32</td>
<td>2</td>
</tr>
<tr>
<td>Frequency</td>
<td>1.91</td>
<td>1</td>
</tr>
<tr>
<td>IMLSDCC search queries:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>1.94</td>
<td>2</td>
</tr>
<tr>
<td>Frequency</td>
<td>1.60</td>
<td>1</td>
</tr>
</tbody>
</table>

**Search Options**

Most of the search queries in both digital libraries were basic keyword searches. The use of one or more advanced search options was observed in slightly under 15% of OH search queries and in almost 20% of IMLSDCC search queries overall.

<table>
<thead>
<tr>
<th>Advanced search</th>
<th>OH (% of queries)</th>
<th>IMLSDCC (% of queries)</th>
</tr>
</thead>
<tbody>
<tr>
<td>search limit</td>
<td>12.58%</td>
<td>13.37%</td>
</tr>
<tr>
<td>fielded search</td>
<td>0.61%</td>
<td>5.00%</td>
</tr>
<tr>
<td>phrase search</td>
<td>1.32%</td>
<td>1.75%</td>
</tr>
</tbody>
</table>
As part of advanced search, both digital libraries provide options to limit collection-level search results by the type of objects in digital collections (for example, to retrieve only collections that contain photographs), and to limit item-level search results to a specific digital collection or a group of collections. Over 12% of search queries in OH and 13.37% in the IMLSDCC included one or more of these search limits (Table 2).

Fielded search option — advanced search where the user is allowed to search either by author or by title and subject words — is available for item-level search in both digital libraries. Fielded search was observed in 0.61% of OH and 5.00% in the IMLSDCC search queries. Use of another advanced search feature — phrase search with quotes — was observed in 1.32% of search queries in OH and 1.75% in the IMLSDCC.

In both digital libraries, collection-level search was used more often than item-level search option, although the latter is less prominently located on the page. In the OH, collection-level search accounted for 52.52% of all unique search queries; in IMLSDCC this proportion was even higher (57.21%).

Search Categories

In this subsection, results are presented as percentages of “known queries” — the queries that the authors of this paper were able to categorize into one or more search categories. There were a total of 2,599 such queries in OH sample, and 3,039 in IMLSDCC sample. As shown in Figure 1, the top two user search categories in the domain-specific OH were place (e.g., “Wyoming”) with 34% of queries and object (e.g., “quilt”) with 31% of queries. It is worth noting that these categories belong to FRBR Group 3 of entities, or subject entities. Another Group 3 search category — concept (e.g., “public housing”) — was the fourth most often occurring search category (17%). However, the fourth FRBR Group 3 subject entity — event (e.g., “1935 meat strike”) — was observed much less than the other three (10%). The FRBR Group 2 search categories — person (e.g., “James Henry Blake”) and corporate body (e.g. “United Fruit Company”) — were observed in 26% and 14% of queries respectively. The work search category (e.g., “Crafte of conjureynge and howe to rule the ffierye spiritts of ye”) was observed in 9% of queries, while the class of persons (e.g., “pilot”) and ethnic group (e.g., “Navajo”) search categories occurred in 8% and 5% of queries respectively. Finally, family search category (e.g., “Wright brothers”) was observed in only 0.43% of unique search queries.

Similarly, the top two user search categories in the domain-independent IMLSDCC were place (32% of queries) and object (31%). Concept was the third most frequently occurring search category in IMLSDCC with 27% of search queries, considerably more than in OH. The event category occurred much less often than other three FRBR Group 3 search categories and somewhat less often than in OH (7%).
Person and corporate body categories were also found less often than in OH: in 19% and 9% of search queries. The work search category was observed in the same proportion of queries as in OH (9%). Class of persons and ethnic group categories occurred in IMLSDCC considerably less frequently than in OH: in 5% and 2% of queries respectively. Finally, family category was observed in only 0.3% of queries, which is somewhat lower percentage than in OH.

Discussion of Findings

Preliminary findings of this exploratory comparative study of user searching in domain-specific and domain-independent large-scale digital libraries reveal several notable differences and some similarities as well. Among the two digital libraries with identical interfaces, in domain-independent digital library advanced search was used somewhat more often. While the rate of use of search limits was comparable, the use of phrase search differed substantially, with higher numbers for domain-independent digital library, and the use of fielded advanced search was significantly higher in domain-independent digital library. The proportion of collection-level searching was high for both digital libraries but also somewhat higher in domain-independent digital library.

Overall, the level of advanced searching observed in both domain-specific and domain-independent digital libraries is high compared with the findings of studies of user searching on the web (e.g., Spink & Jansen, 2004) or in online databases (e.g., Nicholas et al., 2009). This may indicate higher proportion of domain expert users in large-scale digital libraries, as a number of user studies (e.g., Hembrooke et al., 2005; Wildemuth, 2003; Zhang, Anghelescu, & Yuan, 2005) report that selection of advanced search options increases with increase in the user domain knowledge. However, the actual level of advanced searching in IMLSDCC and OH is still significantly lower than the level of preference (81%) for advanced search that was observed in a survey of the users of a similar large-scale digital library (European Digital Library) conducted by Agosti and colleagues (2007).

Both average length and average frequency of search queries were substantially higher in domain-specific digital library. The authors of this note are aware of only one previously published study (Jones et al., 2000) that measured search query lengths in a comparable, large-scale digital library environment. Current study's findings for domain-specific digital library query lengths are consistent with Jones et al. results, but domain-independent digital library search queries are found to be shorter.

Results of this study mostly agree with Bates’ (1996) taxonomy of key search query types. Search queries in both domain-specific OH and domain-independent IMLSDCC most often include person, place, and concept search categories, which correspond to Bates’ query types: names of individuals, geographical names, and discipline terms. However, event category, which corresponds to Bates’ key search query type of chronological terms, is found significantly less often in both digital libraries. This study has also revealed additional key search query category — object.

The same 4 search categories — place, object, person, and concept — occur more often than others in both domain-specific and domain-independent digital libraries. However, 5 search categories (person, corporate body, event, class of persons, and ethnic group), are used substantially more often in domain-specific digital library while one search category (concept) is used significantly more often in domain-independent digital library.

Finally, this study revealed the overall prevalence of subject searching in large-scale digital libraries. Three out of four most frequently occurring categories in both types of libraries — place, object, and concept — match Group 3 (subject) entities in the model of Functional Requirements for Bibliographic Records (IFLA, 1998; 2008) and therefore are in fact subject search categories.

Conclusion

This study provides empirical data to support digital library developers’ decisions regarding information organization in large-scale digital libraries. For example, based on the findings of this study, it would be wise for developers of both digital libraries in the domain of US history and domain-independent libraries to prioritize recording places, objects, persons, and concepts in their metadata records. In metadata creation for digital libraries in the domain of US history, special attention should be given to documenting corporate and personal names, dates, classes of persons and ethnic groups. At the same
time, developers of domain-independent digital libraries need to ensure recording concepts, places, and ethnic groups in metadata, and providing advanced search options, especially for fielded search.

Overall, user experience can be improved if large-scale digital libraries — both domain-specific and domain-independent — supply an option to limit search results by geographical area, which is suggested by the high proportion of place searching observed in this study. Last but not least, the prevalence of subject searching among the users of both domain-independent and domain-specific digital libraries, which was observed in this study, suggests that provision of the subject-based advanced search option should be prioritized in the design of large-scale digital libraries, regardless of domain.

While accurately and unobtrusively capturing objective data on the actions of digital library users, transaction log analysis is unable to provide data on users’ motivations and reasoning behind these actions. Triangulation of transaction log analysis results with results obtained through other methods, for example through think-aloud protocol observations, would overcome this limitation and help develop a fuller picture of user searching in domain-specific and domain-independent large-scale digital libraries.

References


Utilizing Public Displays to Enable Location-Centric Information Sharing

Nicholas Wittison
Information & Interaction Research Group
School of Computing & Information Systems
University of Tasmania
nic.wittison@gmail.com

Christopher Lueg
Information & Interaction Research Group
School of Computing & Information Systems
University of Tasmania
christopher.lueg@utas.edu.au

Abstract

Location-centric information sharing means that information can be shared at a specific geographic location by any person, regardless of the person's actual geographic location. Conceptually, we position location-centric information sharing as a contribution to P3 (People to People to Geographic Place) research which is about understanding the (new) meaning of place in social networks. Technically, we describe a proof of concept that links recent developments in mobile technology (smartphones and tablet computers, such as iPads) and large public displays and that demonstrates how off-the-shelf technology can be used to enable the remote sourcing of highly location-sensitive information. We are currently looking into installing the AskAway prototype for long-term embedded evaluation at a number of locations including universities, start-up companies and coffee shops.

Keywords: location-centric information sharing, information access, human computer interaction, mobility, public displays, situated displays, locative media

Introduction

"Push" and "pull" are two fundamental ways of accessing information. In a nutshell, pull means that the user initiates the information transaction whereas in the case of push the transaction is initiated by the system. "Information at your fingertips" (IAYF) can be regarded as an instantiation of "pull" since it is the user that would determine what information he or she wants to access, and more importantly, where (typically the user's location) and on what output device (typically the user's device, but not necessarily so as we will show below).

"Push" has become omnipresent in the age of smartphones notifying the user of incoming messages with the transaction often being initiated by another user. Push allows to specify the recipient but not normally the recipient's location. Following the IAYF vision it would not be necessary to specify location because of the recipient being in control of the where and how the information is received. There are situations, however, where pushing information to a specific location as opposed to a specific user may be desirable.

Consider the following scenario of Courtney, a backpacker who is seeking out the best gigs in town:

Courtney is a backpacker from Canada who loves travelling, surfing and listening to live music with a preference for alternative and grunge. Planning to head North from Sydney, Australia, she already knows that over the next few weeks, she'll be in Sydney of course, the home of legendary rock band Midnight Oil, Newcastle (Silverchair's hometown), and Brisbane (home of Powderfinger). A few weeks later she'll arrive in Cairns in Australia's tropical North.

Searching the internet brings up a few local community websites detailing the attractions of the towns she will pass through but not the specific gig information that she is looking for. As seasoned traveler Courtney knows that often, upcoming gigs are promoted on posters around pubs and backpacker hostels. Plus... backpackers often know a great deal about upcoming gigs!
When Courtney comes across a web site "AskAway" that allows to push information requests to specific locations she decides to give it a go. She uses her Facebook account to login and posts questions about upcoming gigs to a number of "AskAway" message boards that have been installed in backpackers around the country.

To her surprise, within a few hours she knows about Sandringham and Glebe Hotel gigs not to miss when in Sydney [posted by Sharon from England who has been in Sydney for the past couple of weeks, and Bruce from the US]; she even learns that the iconic Hopetoun Hotel is closed [Aaron from Scotland learned it the hard way when standing in front of closed doors]. Gigs in Newcastle look promising too whereas it seems to be a bit early to inquire about gigs further to the North.

The scenario helps explain what we mean by location-centric information sharing: that information can be shared at a specific geographic location by any person, regardless of the person's actual geographic location. Courtney, for example, is asking questions at locations that she intends to be at in the near future. In a way, location-centric information is the opposite of ubiquitous information access or IAYF where the specifics of location or hardware are abstracted away. The IAYF vision has been promoted for decades (see e.g., McFedries 2009) even though there are open questions about the specific representations of information that may be required in certain situation, suggesting that IAYF may not always meet people’s needs (e.g., Perry et al., 2001).

Related Work

Linking information and location has been investigated by a range of disciplines. Work in mobile and ubiquitous computing looked at using location to trigger alerts when the user (or, more specifically, the user's device) is at a specific location (e.g., Beigl 2000).

Locative media (e.g. Bilandzic & Foth 2012) is a term used for information that is "deposited" at certain locations and subsequently made available only to users that are physically co-located. It is a matter of implementation whether information is only available at a location because of physical transmission constraints (e.g., Maunder et al., 2007 using range limited BlueTooth) or whether the user's location is determined to check if the user is co-located, e.g., by using IP address or GPS like the 'National Mall' iPhone app.

The idea of leaving behind information at certain locations has been explored in a range of systems including 'GeoNotes' (Espinoza et al., 2001) and 'ActiveCampus Explorer' (Griswold et al., 2003), both allowing users to contribute content related to activities or events that take place there. We consider electronic 'graffiti' as implemented by GeoNotes an early example of locative media since depositing and accessing information left behind was constrained in specific ways: users were presented with content that previous visitors of the physical space left behind only if they were at the same geographical location where the information had been deposited.

Sharing of information at a specific location was also investigated by Annay & Strohecker (2009) using an installation of a public opinion forum ‘TexTales’. TexTales utilized a large screen projection to display user contributed photos. Bystanders were encouraged to annotate the photos via SMS in order to nurture collaborative public expression and to demonstrate the potential of intermodal conversations in urban spaces. Similarly, ‘Discussions in Space’ (Schroeter & Foth 2009) aims to engage community members co-located in a physical space by proposing topics of discussion on an embedded screen and allowing users to contribute via SMS or Twitter.

The ‘nnub’ system by Redhead et al., (2010) also aims to connect community members and help them share information from within their community. A large touch screen display allows for users to read and contribute to the notice board style forum. The ‘nnub’ system is undirected, like a traditional noticeboard, and discussion can focus to any particular topics that are raised. Since 'nnub' tries to cater for a wide variety of topics, it can suffer from the same information overload and clutter issues presented by traditional notice boards or web forums.

‘CityFlocks’ by Bilandzic et al. (2008) is an unusual system in the sense that it links visitors to a database of user contributed reviews about the physical location. Furthermore it also links visitors directly to people inhabiting the location who volunteered to be contacted in an attempt to measure the preferred method of accessing this information.

The latest iteration of 'Brisbane Hot Spots' called 'YourScreen' (Seeburger & Foth 2012) allows people to push digital media to remote screens installed at bus shelters via Twitter but does not yet allow for bi-directional information sharing as required in the aforementioned scenario.
To sum up, the notion of location-centric information sharing that we depicted in the scenario is different from aforementioned works where information is 'pre-installed' at certain locations to suit specific purposes as is the case with locative media. Location-centric information sharing is also distinct from targeted but unconstrained access to information from the internet or similar information sources even if the information is related to the location in a spatio-temporal sense.

Specifically, public-display based location-centric information sharing as described in this paper is a contribution to P3 (People to People to Geographic Place) related research which is about understanding the (new) meaning of place in social networks. It used to be the case that place-based ties were formed around specific geographical locations, such as villages, whereas ICT based ties can be formed and maintained over distance which means respective communities are largely independent of geographic locations (e.g., Jones & Grandhi 2005; Rainie & Wellman 2012).

**Enabling Location-Centric Information Sharing**

In the previous section we pointed out that conceptually, location-centric information is different from the related work that we discussed since none of those approaches support an unconstrained, location-specific 'push' operation that is required to enable location-centric information sharing. In this section, we discuss AskAway which is a prototypical implementation of location-centric information sharing. We briefly describe the client-server architecture as well as how users interact with AskAway using common social media tools and platforms including Facebook and Twitter.

Before discussing the actual implementation (see Wittison 2011 for details) it is worth noting that regarding user acceptance of P3 systems, Jones et al. (2008) investigated the need for P3 systems and what the general public would want and be willing to contribute to them if they were around. Out of 500 people surveyed in Manhattan, 54% wanted to read place linked community generated comments, and over 45% of people are willing to share personal location information with others to enable P3-System services.

Relevant findings from four experiments conducted by Jones et al. include that a) information needs about a certain location were closely tied to the activity subjects were doing or planning to be doing at that location. Whether or not the information was static or dynamic made a difference depending on the location (e.g., a static train timetable vs. a dynamic time waiting to be seated at a popular restaurant). A key insight in the context of this paper is that the more dynamic the information was the greater need for some way of obtaining it.

The findings by Jones et al. support the backpacker scenario depicted earlier in the paper in the sense that a) there is a need for dynamic access to specific information, and b) there is a need to use this for both planning and current activities. Another key finding is the confirmation that people would be willing to share this kind of information with others.

**Interaction Model**

Typically, AskAway screens are installed in safe locations such as in pubs and backpackers (hostels). In order to determine how information would be displayed and accessed on remote AskAway screens it is necessary to specify how users will interact with the system. If, for example, users are to interact with the display directly, then screen real estate must be reserved for an on-screen keyboard if users are to write on a virtual keyboard. That kind of interaction is fashionable but put limits on the placement of the screen as it must be placed within physical reach of users and it also increases the risk of the surface being damaged.

Interaction with large screens using mobile phones has been researched in a number of projects including Lyle et al.'s (2012) describing interaction with a 'democratic jukebox'. A relevant finding is that the social setting may influence how a user wishes to hold their mobile device when participating.

Tang et al. (2008) found that supporting covert engagement and interaction was a way to help with people crossing the threshold from viewing to participating. Covert engagement and interaction is the process of allowing users to interact and participate in a system without being easily identified as doing so, e.g. interacting by the users personal phone as opposed to interacting via a central kiosk. The particular method used by Tang et al. (2008) was SMS. It was found that this method lets people interact with less fear of social embarrassment. People who were somewhat anonymous when interacting via their mobiles also left longer and more thoughtful messages. We also considered Floyd et al. (2007)
arguing that embedding a new functionality into what potential users do every day will create an easier-to-use system.

We therefore decided to leverage the ubiquity of social networking services and utilize both Twitter and Facebook as ways for users to respond to questions shown on the AskAway display. Users of these services will know intuitively how to contribute to AskAway as the process for submitting answers is the same as replying to a tweet or to a posting on a Facebook wall, both of which are actions that users of these services would have experienced before. The decision to use Twitter and Facebook to store answers also removes the need for specialty infrastructure to support AskAway since all the answers are accessible from the web.

**Large Shared Display Application**

The large shared display application for AskAway was developed for Apple’s iOS 5 iPad platform (with physical screens linked to the iPads). Several choices were analyzed including custom designed systems, desktop computer platforms and various mobile tablets.

The newly released Apple iPad 2 was eventually chosen because it possessed all the features required, including computing power, and also possessed the ability to mirror to any AirPlay enabled screen over WiFi.

**Large Shared Display Interface**

AskAway’s interface for the large shared display application (see figure 1 below) requires two screens. The first connects the application to Twitter and Facebook, which requires usernames and password logins at some point. The second screen is the main display. The main display is made up of 4 sections which have been highlighted in the figure below showing:

- a static section that describes how onlookers can participate in answering questions by using their Twitter or Facebook accounts,
- past questions cycling on the display, and the number thereof, as well as the most recent question being asked,
- answers for on-screen questions posted using Twitter, and
- answers for on-screen questions posted using Facebook

**Website Plugin**

A website plugin is the companion to the large shared display application described above. The plugin is situated on a community or business website and allows for people to push questions to remote AskAway screens. Plugins can be customized such that they only submit questions to the displays that they are associated with.
Discussion

The primary objective of this research is to investigate how off-the-shelf technologies could be used to enable location-centric information sharing by allowing users to push information requests to specific locations and pull any answers to respective inquiries. In our case the information is shared on public displays installed at remote locations that user can interact with using their Twitter and Facebook accounts.

This work makes two distinct contributions. Conceptually, location-centric information sharing provides a contribution to P3 (People to People to Geographic Place) research which is about understanding the (new) meaning of place in social networks. Location is at the core of P3 research but as yet the P3 framework does not provide the conceptual means for location-centric information sharing as explored in this paper.

Jones et al. (2004) do mention the capability of the 'Active Badge' system (Want & Hopper, 1992; Want et al., 1992), one of the earliest P3-Systems, of routing communication to a specific location: "Active Badges were designed to facilitate communication. For example, phone calls could be routed to the phone nearest to an individual based on the last reported location" but in the case of Active Badge, routing communication to a location is merely a byproduct of routing a call to a person; neither Want et al. (1992) nor Jones et al. (2004) discuss the potential of using Active Badge-like P3 systems to facilitate communication at and about a specific location which is what AskAway facilitates.

Technically, we described a proof of concept that links recent developments in mobile technology (smartphones and tablet computers, such as iPads) and large public displays and that demonstrates how off-the-shelf technology can be used to enable the remote sourcing of highly location-sensitive information.

We are currently looking into deploying a number of AskAway installations at locations around town including several campuses of the local university, start-up companies and, most importantly, coffee shops for long-term embedded evaluation (e.g. Churchill et al. 2003).
References


Context and Collection:  
A Research Agenda for Small Data

Amelia Abreu  
University of Washington  
ameliacabreu@gmail.com

Amelia Acker  
University of California, Los Angeles  
aacker@ucla.edu

Abstract

In recent years, big data has become a prevalent issue for information studies. In an era of big data, can we contemplate research data that relies more on the context of creation than volume and variety of source. In this note, we report on early findings of phenomena we identify as small data. Despite the outpouring of critique and theoretical assertions related to big data, little attention has been paid to the collections, researchers and collecting institutions that get left out the rhetoric of big data. We present criteria for small data and explore some of the issues inherent in developing small data research. The resulting analysis develops future directions towards a comprehensive small data research agenda. We also develop and discuss factors for consideration in context, preservation and access of both big and small data.

Keywords: Big data, data curation, infrastructure studies, small data

Introduction: What Big Data Leaves Behind

In recent years, claims of a big data revolution have generated significant response in information studies. Big data has been positioned as both a must-have commodity and resource in business, government, academic research, and military applications [1–3]. Oracle, Microsoft, and Intel have each developed big data analytic tools for enterprise [4–6]. Likewise, academic big data initiatives in disciplines such physics and astronomy have promised new levels of analytics and discovery. Even the National Endowment for the Humanities has created the “Digging into Data Challenge,” partnering with international collaborators to challenge social scientists and humanists to incorporate large data corpora and data intensive computational techniques into their research agendas [7], [8]. We have the opportunity to consider context, value and research methods now that the turn towards big data has been identified.

In this paper, we define the phenomena of small data as it relates to big data narratives in information science and social science research. Critics have identified how big data will create new digital divides in analysis of and access to data, as well as tools and levels of expertise that can be applied once it has been gathered [9], [10]. How do new contexts of big data subsume the possibilities for rich and faceted interpretation of data that is “too small” to fit in the current discourse of processing power, distributed computing resources, walled platforms, and data analytics?

Small data, as we refer to it, exists as a growing area of study that has been overlooked in the big data ecosystem. By engaging with small data, we may critique the assumptions, processes and commitments of big data. Because big data is seen as an extension of the data-intensive, fourth paradigm of science, it provides information scholars the opportunity to look closely at the places where small data lives, is used for discovery, and has been preserved for access in decades past.

In this note, we propose the following questions for small data research in the era of big data:

• What is “small data” and how does it differ from big data?
• How can the value of small data research be articulated in response to the drive for big data?
• How can we design tools and information systems for small data?
Related Work

In developing a small data research agenda, we draw on two main areas of work: critiques of big data research and possible tactics for small data approaches, located in the study of infrastructure, the digital curation movement and personal digital archiving.

Critiques of Big Data

While business, scientific and social research are using big data to drive change and ask new questions, some critics have expressed shortcomings. The most common critique of big data has been the question of access to data and in turn the possibilities for analysis. Manovich has coined these access issues as a “data analysis divide” [10]. One main point of contention is that of agency in relation to personal information. As users contribute content and data in social media platforms, it remains to be seen who gets access to these data sets and under what auspices. The majority of big data research takes place in the environments with the most resources, either in corporate settings or top-tier research universities. Further students and faculty from such universities are more likely to gain access to corporate research settings and granting agencies, and thus “set[ting] up new hierarchies about ‘who can read the numbers’” [9].

Another problem with big data applications is the range of the possible social scientific questions that can be asked of it in combination with scale, or what some communication scholars have called ‘internet time’ [12–14]. Big data sources collected online are often incomplete because of closed-platforms, privacy policies, speed of change, and limited information access. Many of the research methods that are being applied to big data are untested. David Karpf has written about some of these methodological cul-de-sacs that are results of “endemic problems associated with online data quality,” noting that researchers who work with big data are “well aware of its limitations. Spambots, commercial incentives, proprietary firewalls, and noisy indicators all create serious challenges.” [12].

Big data are not isolated data sets, cordoned information, subject or platform specific data. As danah boyd and Kate Crawford argue, the most important aspect of big data is the possibilities for relationality with other data: “the value is in making connections between sets” [16]. Corporations have developed big data management services for scraping and cleaning unstandardized data for storage, access, and analysis [4–6]. Relating data sets is only possible if the owners, creators, and generators/ producers of data have the means to do so.

Infrastructure studies, Data Curation, and Personal Digital Archiving

The convergence of big data practice and intensive-data discovery in scholarship and professional practice in information science can be seen in the areas of infrastructure studies, the recent history of data curation and the emerging field of personal digital archiving (PDA). Geoffrey Bowker and S. Leigh Star have written extensively about the value-laden and performative aspects of infrastructuring that are involved when heterogeneous data-sets are combined in data-intense science [17–19]. By analyzing how values are ascribed in the structure and layering of technology, protocols and standards, we may locate the performative nature of databases and our abilities to access that data in different ways. We can locate how big data enables new ways of knowing by employing tactics from infrastructure studies [20]. Additionally, by being attuned to change across information infrastructures we may also locate how we have known in the past, and further the ways of knowing with small data before and during the big data era.

Data curation techniques focus upon the quality, trustworthiness, re-use of data for discovery, and they are often framed as confronting the ‘data deluge’ [21–23]. Small data research agendas can build upon data curation techniques that rely upon the continuous enrichment of data; including, starting from the contexts of creation, a commitment to metadata and a deep understanding of its shifting, often ephemeral qualities [24], [25].

Another area ripe for small data scholarship can be found in studies that take seriously the personal digital archiving practices of individuals documenting their digital lives [26–28]. While description
and expectations for future use are recurring themes in narratives of big data, they also have long history in archival studies and practitioner research focused on enduring value and stewardship [29], [30]. Engagement with personal digital archiving is another means of getting at the nuances and variety of possible small data research questions and applications.

While the “big data rich” can be found in industry, government, and research universities, the small data rich may be all around us, in the cultural heritage institutions like libraries, museums and archives that are committed to public access and accountability [9], [31]. Moreover, humanistic, historical and social science methods emphasize bounding research data that privileges context, the process of creation and capture, and emphasizes principles of preservation.

Researching small data begins by countering big data’s hype. The study of infrastructure, the expertise and analytics of data curation, as well as the turn towards personal records management in PDA each confront the nuances, half-life, affects and influences of the big data era.

**Defining Small Data**

In order to define small data, we must first articulate its distinctions. In this section, we posit six central distinctions of small data: motivation, data collection, context, affect, archival engagement, and retention. By doing so, we develop an operational set of factors for analysis and discussion (Figure 1.).

<table>
<thead>
<tr>
<th>Small Data</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data collected purposefully in defined settings</td>
<td>• Data collected transactionally in a variety of settings, across platforms</td>
</tr>
<tr>
<td>• Data collection is local: conducted by individuals or teams</td>
<td>• Data collection is automated: conducted by retrieval tools</td>
</tr>
<tr>
<td>• Data collection is held to disciplinary and community standards</td>
<td>• Data collection not standardized: dependent on setting, platform, and terms of service</td>
</tr>
<tr>
<td>• Data set is part of research archive and made visible to community</td>
<td>• Data set is proprietary, access limits</td>
</tr>
<tr>
<td>• Institutional, professional, and disciplinary standards for rigor and ethics in data collections</td>
<td>• Lack of standards for rigor, quality and ethics in data collection</td>
</tr>
<tr>
<td>• Human labor in data collection as professional or scholarly activity</td>
<td>• Data collection is de-professionalized labor</td>
</tr>
<tr>
<td>• Informed consent of subjects</td>
<td>• Implied consent of subjects</td>
</tr>
</tbody>
</table>

Figure 1. Key Distinctions between Big and Small Data.

To begin, small data is collected differently than big data, usually with defined parameters and boundaries. Instead of collecting data indiscriminately and automatically, small data collecting motivations are articulated at the outset. These motivations are generally purposeful and to a certain degree, conscientious, stemming from a research question, a hypothesis, or an individual or group mandate.

Moreover, the professional, scholarly, or cultural mandate to collect data holds the work to a higher standard than that of big data collection, with considerations for subjects, context, and impact of data collection reflecting an established ethic and rigor. In Jenna Burrell’s example of ethnographic data, the research data set documents “some kinds of things straightforwardly, but not others” [15]. In scholarly communities such as ethnography, data collection is professional work of utmost importance.

**Context** is perhaps the most important factor that distinguishes the small data approach, as it is the most difficult factor to regain when lost. For ethnomusicologists, fieldwork and data collection “should
happen where music happens” [32], and thus an ethnomusicologist’s data set consists not only of music, but other observations as to its context and “happening”. In this community, big data methods “can at once benefit our work and magnify our reflexive anxieties about the impact of our data collection on our ethnographic integrity” [32].

Small data collection requires a deeper affective engagement, both in terms of professional labor and personal commitment. For ethnomusicologists and other scholars interested in the contextual and affective qualities of their research data (such as art historians or literary scholars), existing big data tools cannot parse, analyze, or comprehend the significance of their data sets with the level of critical rigor these types of inquiry demand. Although big data research has begun to venture into affect inquiry [33], the resulting research is dramatically different than that of traditional scholarship in the field [34].

In the humanistic tradition, use of data as texts, artifacts, or observations is seen as an act of archival engagement, one deeply ingrained in the values and ethics of scholarship. Literary scholar Lauren Berlant recounts a colleague’s comment, “I hate your archive,” in relation to the texts and films she presented for analysis in a talk. Berlant asks, “Was this an aggressive disciplinary question?” pointing out the affective dimensions selection, collection and interpretation of such data for humanists [35].

Small data research demands not only critical attention towards the collection process, but also for standards of retention. Across disciplinary lines and research settings, scholars have established different retention standards to ensure quality, respect privacy, and provide access to their data/research process. For example, historians cite publicly available records; while anthropologists are expected to anonymize field notes; while poets use manuscript drafts as part of a literary archive [36]. In big data research, retention and custody of data sets remains an unsolved and often controversial issue [37].

Small Data Futures

It is highly likely that small data will intersect with big data in the coming years, and that small data and big data research will overlap in complementary praxis. Hybrid models of integrating big data with small data have emerged, such as Anderson et al’s work [38] on integrating sensor data with ethnographic fieldwork, and Batty’s [39] agenda for human geography data. For many researchers seeking to integrate big data methods, a key challenge is making big data smaller: scaling available data to the parameters of the focus of research questions. As we go forward, we anticipate future research in both design for small data and policy for its access and preservation.

As we have argued, small data is remarkably dependent on context, the future of small data research must be innovative in this regard. The contributions of infrastructure studies, data curation and personal digital archives can aid in developing a more nuanced small data model. While some may argue that big data will usher in an “end of theory” [41] scholars such as Nigel Thrift [42] and Donna Haraway [43] offer valuable theoretical contributions for framing personal and social phenomena in the big data environment. Indeed, the persistence of small data shows that life in the era of big data is complex, but that individual agency is both possible and necessary.

References


“I’ll Be Watching You”:
What do Canadians Know About Video Surveillance and Privacy?

Brenda McPhail
Faculty of Information/U Toronto
Brenda.Mcphail@utoronto.ca

Joseph Ferenbok
Faculty of Information/U Toronto
Joseph.Ferenbok@utoronto.ca

Roxanna Dehghan
Faculty of Information/U Toronto
Roxanna.Dehghan@utoronto.ca

Andrew Clement
Faculty of Information/U Toronto
Andrew.Clement@utoronto.ca

Abstract

The growth of video surveillance systems and their augmentation by biometric and smart algorithms has significant implications for personal privacy. The growth of ICT networks and technologies such as face recognition make it increasingly important for people to know their personal information rights. Our project assesses what citizens know about the privacy implications of developing video surveillance technologies. We have three key questions: 1. What do Canadians know about their visual (video) information privacy rights?; 2. Does information about the current technologies and capabilities change people’s perspectives on visual privacy policy?; and 3. What information do people need? Three research stages are planned: interviews regarding video surveillance, workshops and round-table discussions on visual information policy, and a Public Forum to promote discussion among stakeholders including citizens, government agencies, vendors, academics, civil society, and media. This Research Note summarizes the project and reports on preliminary findings from 126 interview participants.

Keywords: information policy, privacy

Introduction

Every breath you take
Every move you make
Every bond you break
Every step you take I’ll be watching you

--The Police

Visual information collection practices in publicly accessible spaces affect us all. However, it would seem that there is surprisingly little public information or discussion about where, when, or even whether such information should be collected, and how it can be utilized. Perhaps due to the success of the dominant “public safety” rhetoric associated with the use of camera monitoring in both private and public spaces, there is generally relatively little outcry or concern in situations where cameras are installed and monitoring introduced. Public perception of video surveillance is primarily shaped by the grainy pictures of wanted criminals released on the evening news, and by stories crediting such pictures with helping to identify those individuals, leading to their arrests. In reality, research suggests that the effectiveness of video surveillance for increasing security is uncertain, at best, but up to this point, such research has had relatively little effect on use patterns or public perception (Greenberg & Hier, 2009).
Mainly in response to the pressures for enhancing security and the increased availability of inexpensive, interoperable digital hardware and software, cameras have become commonplace. Furthermore, the capabilities of video surveillance technologies have grown and changed over the past decade, opening up the possibility that the visual information collected might soon serve additional purposes. Rapid development in video analytics means that existing infrastructures may be incrementally upgraded to have profound implications for the mining and collection of visual data.

As the Office of the Privacy Commissioner of Canada (OPC) notes in their Guidelines for Overt Video Surveillance (2008), individuals have “the right to lead their lives free from scrutiny”. However, this right may be increasingly compromised as businesses seek a competitive advantage in their delivery of personalized advertising and individualized services. With the surveillance and computing infrastructure already in place in many spaces, video analytics become an incremental cost with a large potential payoff: to know more about the customer.

Do we give up all rights to our personal images when we walk into a public space? When we enter a store, do we by default consent to allow the store to examine our movements minutely in order to assess the effectiveness of their sales displays or our relative interest in competing products based on what we look at longest? What level of privacy do individuals expect in public spaces, or in public yet privately owned spaces, and what should they expect? Do their expectations conform to reality? We are at a critical juncture in the development and deployment of these technologies, as their use is relatively new and policies, procedures, and uses for the information that is visually captured for analysis, are just evolving. This period of change and flux is the ideal time for the development and deployment of privacy-protective guidelines for the use of these technologies. And we would argue that for these guidelines to be responsive to the concerns of citizens, it is necessary to determine what those concerns are. Our research begins to do this, as we look at what these changes in the capacity and use of video surveillance mean from the perspective of the surveilled subjects in terms of their well-being and privacy rights.

Our non-traditional project, based in Canada and funded by the OPC, combines research and public education goals. It extends work on understanding the relationships, attitudes, policies and practices of stakeholders in the video surveillance techno-social network, but in its final stages will also seek to provide some guidance to the OPC in developing materials that address specific questions or concerns that people may have about video surveillance practice or policy in the Canadian context. While our work is situated in Canada, the overarching policy and ethical issues around camera surveillance reach beyond borders.

Background

Much of the research into the use and effectiveness of video surveillance, and attitudes towards it comes from the UK, where CCTV cameras are ubiquitous (Norris & Armstrong, 1999). However, recently the acknowledged ineffectiveness of CCTV surveillance at reducing violent crime has prompted the current government in the UK to pledge more regulation and a roll back of CCTV cameras (Whitehead & Kirkup, 2011). In Canada the situation appears to be somewhat different (Ferenbok & Clement, 2011). Most cameras have not been placed by publicly supported infrastructures. Instead, the majority of cameras pointed at publicly accessible spaces are commercially run and operated. These commercial operators fall under the Personal Information Protection and Electronic Documents Act (PIPEDA) provisions for informing individuals about the nature of video surveillance, but the preliminary findings of ongoing Canadian research (Clement et al., 2010) suggests that there is little to no compliance with even the most basic guidelines.

This is a particular cause for public concern, because video surveillance technologies are rapidly changing, becoming more powerful, and more easily integrated with ICTs. The drop in the cost of storage and digital sensors, coupled with the promise of algorithmically augmented monitoring means that the days where video surveillance cameras record low resolution images to analogue videotapes is likely numbered. Digital sensors and digital processing promise more than just local passive recording of visual data. Digital signal processing promises to turn image data into actionable information. This potential for real-time intervention for loss prevention or the ability to collect and aggregate large volumes of consumer data is on the horizon. We know that Microsoft’s vision for the future includes ambient identification, with software that can ‘see’ people (Wilhelm, 2011). Armed with object recognition, face tracking, and face recognition, this future suggests a time in the near future where personally identifiable information, like
images of our faces, becomes a portal into our real and electronic lives. Significantly, “this conversion from image data to information continues silently behind the lenses of surveillance equipment generally unseen by the surveillance subject” (Ferenbok & Clement, 2011). The unmitigated growth of video surveillance, in both scope and nature, is encroaching on any form of reasonable control a data subject may have over their information. The use and possible linkage of visual personally identifiable information is an issue that should concern us all.

As researchers concerned with information policy in general and privacy in particular, we assume that these issues are significant, but are average individual Canadians concerned about the information captured by video surveillance? Do people know what may be done with their visual image? Are Canadians aware of their rights in relation to personally identifiable information gathered by video surveillance systems? How much do Canadians understand about video surveillance capabilities? As information researchers and privacy advocates we are working to understand policy implications of new ‘intelligent’ video surveillance, but what we do not know is whether Canadians care. Or more significantly, we do not know whether they have enough information to develop an informed opinion.

We do know, based on previous research, that public opinion surveys conducted by media outlets consistently find high levels of citizen support for camera surveillance (Deisman et al., 2009). But we also know that these surveys are generally conducted after some significant security-related event where cameras have either played a role, or it is supposed they would have played a role had they been present (Leman-Langlois, 2009). More rigorously designed surveys conducted by social scientists tend to find slightly less support for camera surveillance, and show that the location of the surveillance (Leman-Langlois, 2009), the demographic group to which the respondent belongs (Ditton, 2000; Leman-Langlois, 2008), and even the ways in which questions are framed (Ditton, 2000) affect the degree to which people are willing to support camera surveillance. These studies also find that although, on average, about one third of respondents will agree that camera surveillance is a privacy threat, for many of that 33% it is still seen as acceptable because of a perception that it might improve safety or security (Wells, Allard & Wilson, 2006).

What seems clear from the quantitative survey results is that context matters to people when discussing surveillance cameras; what is needed then, is research designed to get information from citizens that is in context. This suggests that a qualitative approach may yield useful results.

**Approach**

To find out what Canadians know about visual surveillance practices and technologies, we need to ask them. The first step of this project, currently underway, is to conduct semi-structured “man and woman on the street” interviews. The questions focus on determining what people know about the video surveillance activities that are capturing their personal visual information, asking what they think about public video surveillance, and determining what they know about their privacy rights regarding such surveillance. We wish understand the extent to which people recognize visual information as personal information, and the degree to which they are informed, or care to be informed, about the practices relating to visual information collected via video monitoring.

Our question sets are based broadly on the OPC’s March 2008 Guidelines for Overt Video Surveillance in the Private Sector, altered to reflect our focus on the surveilled rather than the surveillers:

1. Did you know that this organisation conducts video surveillance?
2. Why do you think this organisation conducts video surveillance?
3. Did you notice any signs or notices that video surveillance is taking place?
4. What do you think the organisation does with the visual images of you and others that it collects? What should they do with them?
5. Does it concern you that you are being recorded? If yes, what are those concerns? If no, why do you feel comfortable with it?
6. Would you ever want to access video information about yourself? How might you try to do that?
7. Do you think video surveillance benefits you as a [shopper/citizen/member of the public]?

To date we have conducted 126 interviews in various urban sites where the presence of private surveillance cameras focused on public space is somewhat obvious. These sites are in the vicinity of major intersections, public squares and community centres. It is our hope that by situating the
conversations with participants in sites where the cameras can be seen and discussed as a concrete presence, people might be encouraged to think about and answer the questions. Data collection and analysis have been iterative and are ongoing.

Subsequent stages of this project will include a series of small group workshops in which items such as examples of signage, either actual or models prototyped in previous research may be used as concrete artifacts that can encourage conversation about what information people want or need to know in order to exercise their privacy rights. The third stage, building on the results of the interview and workshop activities, will be a public information forum, bringing together citizens and experts to explore a range of topics related to visual surveillance and visual information policy. The forum will combine a physical meeting with an interactive Canada-wide internet broadcast.

Initial findings

We are in the early stages of data analysis but have some preliminary findings. Participants range in age from 19 to 65, with education levels from high school up to post-graduate degrees, and work in a variety of occupations including law enforcement, education, child-care, and sales. Interestingly, almost all of our participants state that they are aware that they are often under camera surveillance when going about their daily business in public spaces. Many mention seeing signs informing them that the commercial premises they are about to enter are being monitored by video camera “for [their] safety” or possibly “for theft prevention” although a minority criticized sign location and size--“I have seen small miniscule signs under those set of stairs”-- and/or describe the content as “incredibly vague.” When asked, participants most often say they know that their actions might be caught on a camera when walking through a mall or even down a street. But most state that they do not know is who is really watching, or for what specific purpose(s). Our participants thus far have been largely unaware of the kinds of camera surveillance infrastructure in their local areas, or the ways in which it is, or might be used beyond general perceptions of its use “for security and stuff.” Many do not know whether video surveillance is subject to regulation in Canada, although most assume that it is, but all agree that it is very important that organisations comply with privacy regulations. If asked specifically what they might like to know about camera surveillance and the way it is carried out or regulated, most want to know that surveillance is taking place, “what area is under surveillance” and “what their reasoning was” for conducting the surveillance. There is also concern that the organisations conducting the surveillance use it only for ‘legitimate’ purposes; these purposes are considered to relate to safety and security, with some mention of crime deterrence or prosecution. Participants had varying opinions however about whether video surveillance made them feel safer. Opinions ranged from “What is a camera going to do if I get mugged?” to “I feel indifferent” to “if something is being recorded I feel safe knowing someone will or is able to see what will occur”; answers corresponding to these positions occurred in relatively even proportions amongst the participants. However, all participants asked indicated that they feel it is inappropriate for camera surveillance footage to be used for citizen profiling and the majority dislike the idea of it being used to determine shopping patterns.

Our sample size is still relatively small and fieldwork and analysis is ongoing; thus far our data accords well with studies in other settings, including the US and the UK. Our preliminary findings point to a need for public education around the ways camera surveillance is used, and can be used, and about the current set of laws and regulations that provide citizens with rights to their personal visual information.

Conclusion

Our project combines citizen-perspective focused research and public education in the expanding arena of visual surveillance and consumer privacy. The overarching goal is to find out what people understand about the recent developments in video surveillance associated with the introduction of digital techniques, what concerns they express, and where information gaps exist, in order to inform the development of public education and policy. This research is conducted on the assumption that the potential uses of visual surveillance data need to be made more transparent, and organisations using these technologies need to be more accountable to members of the public in protecting privacy rights.
References


Towards a Method of Documentary Practices for Personal Health Information Management

Matt Willis
Syracuse University
mwillis@syr.edu

Carsten Østerlund
Syracuse University
costerlu@syr.edu

Abstract

Many intellectual fields study patients’ use of health information systems, yet there is little innovation in the variety and richness of methodologies employed to investigate the complex socio-technical relationship between health consumers who use technical systems to engage in the practice of personal health information management. This is, in part, due to the methodological challenges of studying patients’ documentation practices. In this note, we present the early development of a document practice approach to address these gaps. The documentary practice approach is rooted in the socio-technical perspective that social systems and technology are intricately connected and recursive. Documentary practice methodologies are mindful of this entanglement in studying the work practices of actors by connecting practice to the use of documents, both physical and digital, and mapping the patient’s field of documents and technical infrastructure within which health management occurs.

Keywords: practice theory, personal health information management, socio-technical, documentary practice, personal health record

Introduction

Healthcare is an information intensive activity for health consumers, whether they are managing pathology, seeking information, engaging in health prevention activities, or organizing personal health information in a journal or diary. All of these activities are cognitively intense and, most importantly, the methods and strategies a patient can perform are highly individualized and change over time (Agarwal, 2009). In light of the growth of health IT, eHealth initiatives, mobile computing, and personal health records, consumers have a growing list of options when they decide how to engage in health information management. Additionally, the growing number of smartphone and mobile device usage (Smith, 2012), broadband access (Internet Use and Home Broadband Connections, 2012), and the expanding notion of everyday health and quantified self-metrics (Swan, 2009) means that there is more information to track, manage, read, access, visualize, interact and otherwise demarcate. The increase in health information support systems and tools for managing information means that the technical infrastructure (both physical and digital documents) can be configured in a variety of ways. For example, paper plays an important role in personal health information management (PHIM) even when digital tools are available (Piras & Zanutto, 2010). A deeper understanding of how patients and consumers engage in PHIM is a necessary stream of research that will inform the development of health IT artifacts, software systems, facilitate consumer empowerment, and support the core characteristics of patient centered care (Reti, Feldman, Ross, & Safran, 2010). Temporality has been clearly described in previous medical work (Reddy, Dourish, & Pratt, 2006), yet an investigation of patient temporality in relation to PHIM document practices is under-investigated.

Our motivation is grounded in the perspective that the work patients carry out when they interact with personal health information matters and has effects on the delivery of healthcare (Saranto, Brennan, Casey, 2009), as well as the construction of patient driven models of care such as patient centered care.

Copyright is held by the authors.
We focus on making two distinct contributions with this paper. First, we propose a documentary practice approach as a method for addressing current methodological gaps in the literature. Second, we purport that a document practice approach helps expand the common conceptualization of health information management work. Specifically, that PHIM work is not an occasional task individuals decide to practice, but permeates an individual's entire life, akin to illness trajectories (Glaser & Strauss, 1968) or illness careers (May, 2010), and may aid in our understanding of how to best support lifelong personal health records and other PHIM tools.

To better understand the emerging literature of personal health records and how patients harness these opportunities to document health information, we ask the question: How has PHIM been studied methodologically in the literature to date? To answer this question, we gathered a small pool of studies that empirically observe PHIM work practiced with either physical documents, personal health record systems or both.

**Patient Work with Personal Health Information Documents**

The small set of studies we synthesize for methods and findings in PHIM work run from 2005-2010. The literature featured here is selected because of its specific focus on the information management work practices of patients. To begin, Moen and Brennan (2005) utilize interviews, semi-structured questionnaires, artifact collection, and photographs of volunteers. In their analysis, four strategies are developed that explicate the material storage strategies used for health information in homes. Also looking at the role and importance of material objects in organizing a patient’s health information, Palen and Aaløkke (2006) utilize ethnography to observe elderly individuals’ routines for following medicine regimens. They conclude that the participants use their home infrastructure of spatial, temporal relationships to form a home distributed cognition; elders also had computers installed in their homes to document medication intake and other health measures. The next study transitions from studying PHIM in the home to the clinical environment. Acknowledging the ubiquity of health information and the complex ad hoc networks that patients construct, Unruh & Pratt (2008) use an extensive method of in-depth interviews, critical incident interviews, demographic surveys, personal health information collection reviews, photo diaries, log books, and fieldwork to study breast cancer patients’ in situ PHIM work. They develop four patterns of invisible work that individuals encounter in a clinical visit which have implications for both patients’ information work practices and design considerations. These findings underline the difficulty of studying patient information management work, and that to support patient work practices will have beneficial outcomes for patients, providers, and patient driven models of health.

The next three studies take place in 2010, the first of which is Piras and Zanutto’s (2010) semi structured interviews of families’ work with printed health information documents in the household. They observe three record keeping strategies similar to the findings that Moen and Brennan report, as well as four outcomes of work practices that occur in the household. The findings emphasize the flow of documents that patients use and how documents have emotional value to patients, which has an impact on patients’ information management work. Next, Klasnja, Hartzler, Unruh, and Pratt (2010) conduct a qualitative field study of breast cancer patients’ PHIM work in clinical environments. Their findings detail unanchored work, or work that is performed without sufficient support, and identify five patient work strategies and six design principles. In this study the authors showcase the fluidity of information documents and how cognitively taxing and stressful maintaining a patients information ecosystem can be. Lastly, Unruh, Skeels, Civan-Hartzler, & Pratt (2010) utilize ethnography of breast cancer patients in the clinical environment. The findings of this study describe various physical challenges and post visit challenges to patient information work. This work highlights the need to support patient work practices and the importance of information documentation, as work in the clinical environment is stressful and patients’ ability to capture information is fleeting. From this literature synthesis, we move to a discussion of the three themes that emerged from the literature.

**Investigating Patient Documentary Practices**

All of the studies we describe employ qualitative data gathering techniques of either field work or ethnography. While these methods can collect mountains of data, it is often done haphazardly and is rarely linked with good understanding of the larger body of documents and technical infrastructure with which they are connected (Østerlund, Sawyer, & Kaziunas, 2011). A document practice approach builds
on three methodological guidelines that support the collection of any form of information in the field. The first is that all documents are socio-technical artifacts and inseparable from either social or technical structures. When collecting data, this should be acknowledged and lengths should be taken that allow the investigator to connect the document to these originating structures. Second, the purposeful blending of qualitative and quantitative techniques is used in order to analyze documents and document infrastructure with the understanding that they are connected to a larger supportive network. The third guideline is that there is an implicit understanding that documents are a bicooperation of other documents; every document has a sociotechnical infrastructure supporting it. In examining these documents, we can uncover both the embedded knowledge of actors that collaborate around the document, and the technical infrastructure that supports the document (Østerlund, 2007).

Data gathering for the documentation practice approach consists of scoping the relevant network of documents and systems and then gathering as much data as possible, including both physical and digital documentation. For example, if the research site is a hospital environment, the investigators would look at offices, desks, patient computer labs, the medical library, physical and digital filing systems and photo or video documentation as interesting spaces and appropriate tools to capture where, how, and when patients perform PHIM work (Østerlund, Dosa, & Smith, 2005). When collecting documents and details about documents, it is important to pay attention to the organization schemes employed by patients in both physical and digital spaces. This is a method that investigates an actor’s work, in this case a patient, by looking at the documents they create and annotate, and then analyzing those documents for insight into the work practices of personal health information management. Both content analysis and network analysis are encouraged during this phase to elicit the basic structure of a patient’s documentation work.

Once the first step of gathering and mapping documents has concluded, the second step in the document practice approach focuses on tracking the process of documentation work through four data gathering techniques to add depth and understanding to the socio-technical infrastructure. First, it is important to understand how patients organize health information. Do they solely use a personal health record? How do they store this health information? What documents are physical, what documents are digital, and how are they each organized? Do other family members help with PHIM? How much delegation occurs? Who contributes to what document? What documents do healthcare providers see and use? What documents are for patient use? What are highly relevant documents that are opportunistic for patient-provider collaboration? What are the topic and media networks? How does pathology management connect to the use of each document? These are example questions that drive the first data gathering phase. Software can be used to identify which individuals are in the network and track knowledge (Østerlund et al., 2011) across both virtual and physical domains.

The next technique of data collection focuses on linking patients’ document use with behavior over time. Patients can track use of their own PHIM work with pictures, voice notes, smartphone apps, or other participant documentation tools. Researchers can send follow up questions via email along with short interviews. The point of this technique is to track activity of documents and how patients use these documents in a specific place and time. Technique three of data collection is reserved for identifying key documents important to patients’ PHIM work. There should be a consideration of documents that facilitate coordination. Important documents should be tracked in more detail to learn their uses and evolution over time, specifically how they relate to other artifacts and events in the patient’s life. The fourth data collection technique uses interviews and or focus groups with patients to elucidate the meaning that patients associate with specific documents. During interviews and focus groups, the investigator can take this opportunity to focus on constructs and meaning that may be specific to a research project; for example, focusing on how patients’ documents relate to patient-provider clinical visits and help facilitate coordination of care. It should be noted that the most important practice of this method is the reiteration from document tracking to document collection. The described interviews and focus groups should serve as a touchstone to solicit feedback and clarity on the documents that the researcher is collecting, and how those documents link together. From this feedback the researcher can adjust and refine the documents they are collecting and analyzing. We believe this helps create a “sketch” or a window into patients’ documentary practices, reconstructed through the documents they create and interact with. The following diagram was constructed to help elucidate this process. We organize the details of each phase as well as place an emphasis on the reiteration and reexamination of establishing connections between and among documents.
In summation, the following outline details the two steps and four phases of data collection:

I. Initial mapping and scope of documents to be investigated
   a. Create a map of the network of relevant documents. Photos and videos can be helpful here.
   b. Content analysis and network analysis suggested techniques for detailing the structure of document networks.

II. Tracking of document use and recursive data gathering techniques
   a. Understand how each participant organizes documents. Collect logs and employ frequent document queries that can clarify specific uses of a document.
   b. Track participants actions using documents over time, phone calls and follow up emails.
   c. Identify the key documents that are most important to a participant’s work practices. These documents can be tracked in greater detail to focus on specific document networks.
   d. Employ techniques such as participant observation, interviews, or focus groups to establish meaning and understanding of participants’ use of documents. This is a useful opportunity to look at coordination in action.

Future Work

While documentary practices have prior development elsewhere (Geiger & Ribes, 2011; Prior, 2008; Østerlund, 2003, 2004), it is our goal with this note to adapt documentary practices and develop the method into a strong and robust tool to address a clear gap in the literature previously identified. In follow up work to this note we conduct a study of patient personal health information management practices by studying a unique population that utilizes both paper documents and a nationwide electronic personal health record. The goals of the study are in part contained in this document. For example, many of the opening questions are relevant to our study on locating documents for patient-provider collaboration. We also want to stress the importance of scoping, research design decisions, and strategic use of data gathering techniques to make this method manageable and respectful of time, funding, personnel involved, and other real considerations that affect research projects. It is through the follow-up research to this note that we further develop the documentation practice method, providing exemplars of practicality and utility where the method is understood as a modest and mundane regimen of data collection techniques.

Contribution

While each of the empirical studies we describe in the brief literature review touch on some of the steps and techniques detailed here, no study, to the best of our knowledge, of patient PHIM has utilized the documentation practice method in the manner which we begin to detail in this note. Studies that address the infrastructure of PHIM work do so without identifying important documents or fully connecting who collaborates on which documents with whom over time.

In this note, we contribute the identification of a gap in the literature along with a call for the importance of studying PHIM work practices across the affordances of paper and digital documents. We present the use of a documentary practice approach for filling in gaps to systematically and holistically
understand this complex and individualized process of patients tracking, updating, and demarcating their personal health information. We believe that this approach is particularly valuable for work practices that use both physical and electronic documents since it helps synthesize and shine a light on space, place, time, and infrastructure of personal health information management.

References


Smith, A. (2012). Nearly half of American adults are smartphone owners (pp. 1–9). Washington, D.C. Retrieved from Pew Research Center’s Internet & American Life Project


Intelligent Urbanism: Convivial Living in Smart Cities

Stephanie Santoso  
Cornell University  
sms629@cornell.edu

Andreas Kuehn  
Syracuse University  
ankuhn@syr.edu

Abstract

Technology has been hailed as a panacea for global and local challenges in urban development. This research note explores the role technology plays in facilitating intelligent urbanism and considers how Illich’s notion of convivial living can inform urban development. Here, we present one case study on energy that is a work in progress. Two future cases will include transportation and food sourcing. The broader objective of this work is to develop a framework which those involved in the planning, design and deployment of technology in urban spaces can use to integrate conviviality into such environments. This note takes a step in this direction by describing the key elements which can promote convivial tools to facilitate convivial living.

Keywords: smart cities, conviviality, urban development, Ivan Illich, smart grid

“I propose the vision of a convivial society. A convivial society would be the result of social arrangements that guarantee for each member the most ample and free access to the tools of the community and limit this freedom only in favor of another member's equal freedom.”

--Ivan Illich, Tools for Conviviality, 1973

Introduction

Technology has been hailed as a panacea for global and local challenges in urban development. The role of technology is evolving in the face of global challenges faced by cities such as population growth, migration and climate change (United Nations, 2012; Werz & Conley, 2012). This research note explores the relationship between technology and urban development by bringing Illich’s notion of convivial living to this context. Convivial living describes an existence in society that enables a community to choose its own social arrangements. In such a community, individuals have the ability to live their lives in a way that would maximize the use of their creativity and imagination, simultaneously facilitating independence on the part of each person and fostering collaboration between members of society (Illich, 1973). This raises the following questions that we seek to address: (1) To what extent can technology in cities be designed and implemented in a way that facilitates convivial living?; (2) Who should build conviviality into urban spaces through technology?; (3) What role shall residents play in the design and governance of such technologies?

The broader objective is to develop a framework that has two purposes: (1) to examine the extent to which conviviality is present in empirical cases; and (2) to enable those involved in the planning, design and deployment of technology in urban spaces to integrate conviviality into such environments. In its preliminary form, the elements of the framework are theoretically informed by Illich’s work.

The case study discussed here focuses on a smart grid demonstration project originally taking place in Austin and recently expanded to the Dallas-Fort Worth area in Texas. The additional case studies still under development examine the efforts to develop green transportation in San Francisco and urban farming rooftop initiatives in New York City.

Acknowledgements: The authors would like to thank Murali Venkatesh, Gretchen Purser and several anonymous reviewers for their thoughtful insights and helpful comments.


Copyright is held by the authors.
We will briefly define the key terms “intelligent urbanism” and “smart cities,” as they are central to our discussion. Sociologist Louis Wirth asserted that “urban” is not fully or accurately measured by the proportion of total population living in cities (1938). This is because the impact of cities as residential and commercial hubs has radiating social, political and economic effects upon areas outside of city boundaries that are not accounted for when the definition is limited by numbers alone. Accordingly, Wirth believes that urbanism is a way of life that is most often identified with living in a city, but could also be experienced in a non-city area. Thus, urbanism should be thought of as a mode of life that is not limited to the boundaries of the city. Wirth’s perspectives on urban and urbanism will be used within this discussion as it acknowledges that cities and their surrounding areas should be equally considered in examining how technology can be developed and implemented to facilitate a more interconnected and efficient way of urban living.

In this discussion, intelligent urbanism is the leveraging of technology for the purposes of urban development and living. For smart cities, as an anonymous reviewer pointed out, Illich’s notion of conviviality articulates two areas of study: (1) technology in the urban context; and (2) engagement of residents in the creation, implementation and assessment of such technologies. Combining these two areas leads to the establishment of an ecosystem of urban intelligence that includes among others human capital and collective intelligence. There are however other existing definitions of intelligent urbanism. In particular, the Principles of Intelligent Urbanism (PIU) are “a set of axioms, laying down a value-based framework, within which participatory planning can proceed” (Benninger, 2001). These axioms were developed by architect Christopher Benninger in an attempt to address a variety of urban planning issues. There are several tenets which overlap with this work, including “appropriate technology” and “conviviality” (Benninger, 2001).

The definition of “smart city” varies widely but is likely to include cities that are using technology to enhance public infrastructure, such as transportation and utilities, but also private residential housing and the work environment. Smart cities attempt to leverage such resources to address broader scale issues associated with urbanism. Sustainability is a prominent theme within the literature. Herbert Girardet, for instance, calls for the development of sustainable cities, defined as cities which use renewable energy, adopt a circular metabolism and carbon neutral transport systems. In order to accomplish this, technology, in addition to new opportunities in finance, policy and the commercial sector must be utilized. He advocates for the return of the city as a site for convivial living and the establishment of stable communities (Girardet, 1999).

We find that technology can be simultaneously used to constrain and expand boundaries of a city. This means that the concept of urban spaces as local and global becomes more fluid through the development of technology which facilitates the collection, analysis, and dissemination information designed for two conviviality-enabling purposes: (1) Creating technologies which enable individuals to interact with one another and build relationships; (2) Connecting, and coordinating critical activities, such as transportation, utilities (communication, energy, water, sewage) and public safety through a technologically sophisticated system. These two functions of technology in the urban environment map directly to two of Wirth’s most important assertions about urban life. The first is that urbanism overturns and destabilizes the traditional communities within which individuals were once accustomed to living, enabling the individual to live more autonomously (Wirth, 1938). The second assertion is that living in the city increases residents’ sense of isolation (Wirth, 1938). This is considered a downside to urban life and something that technology has and will continue to play a significant role in addressing.

### Convivial Technologies for Urban Development

Ivan Illich criticized the institutions of industrialized society that enslaved individuals through the machines they used at work but outside of their control. With no real control over the mode of production, these individuals became beholden to using specific tools, in specific ways to increase the bottom line. It was this observation which led Illich to assert the importance of humans leading what he called a “convivial life style.” In order to have a convivial life, Illich asserted the need for people to have at their disposal, “convivial tools,” which could range from actual machinery or hand tools, to the skills and education required to operate machinery, to technology.

Illich takes issue with the notion that much of the production technology and machinery are controlled by corporations and other commercial institutions, with very little room left for members of society to use this technology in a way that is self-fulfilling. Because of Illich’s concern of the negative
effects created by purely industrial applications of science and technology, he calls for tools that work “with” people rather than “for” them (1973). Technology as a convivial tool should enable users to maximize their intellectual potential, be creative and live autonomously. Illich notes that “[i]n a convivial society, tools in existence should protect the following three values: survival, justice and self-defined work” (1973). In this sense, technology should be thought of as enabling a higher quality of life for urban residents, rather than limiting their activities to the principle of efficiency or the agenda of certain elite groups and institutions that have power and control over these technologies.

**Case Study on Smart Grid Technologies**

In 2009, an initiative was started that focused on using smart grid technology to more efficiently manage energy supply and demand. This was a collaborative effort among the City of Austin, the University of Texas at Austin and several commercial institutions, which formed the research and development organization, Pecan Street Inc. The resulting smart grid demonstration project is funded by the Department of Energy, the National Science Foundation and corporate partners. Over a five-year period a variety of systems will be tested in residences of the local Mueller community. These systems include: energy storage technologies, smart grid, water and irrigation systems, advanced meters and home energy management systems and new electricity pricing models (Pecan Street Inc., 2012a). The home test locations are the residences of volunteers in the community who allow their energy consumption to be monitored.

Various consumer electronics, automotive, and information technology companies have come on board to develop and test new products and technologies designed to work in conjunction with the smart grid. For example, home services systems integrate consumer smart grid products and services upon a common platform. Commercial partners are developing these service systems, which will address everything from home security and health care monitoring to entertainment and energy management needs (Pecan Street Inc., 2012a). Mueller, the community that serves as the test bed is a redevelopment designed to be a mixed-use, multi-income sustainable neighborhood. Citizen participation was considered a priority in developing this community from the perspective of the City of Austin and the partnering development firm (Mueller Community, 2012). This smart grid demonstration project recently expanded to the Dallas-Fort Worth area in Fall 2012. The information on this case is based on publicly available documents.

While there is clear resident involvement in this project, the extent to which these individuals are able to assume convivial use of the technology is less certain. In this initiative, residents are essentially serving as subjects for an experiment. Their responsibility as volunteers is to use the technologies and devices as they are provided so that commercial and research organizations can collect data on this usage in order to make better decisions about subsequent versions of the products. With this project, there are both significant requirements to be met and limitations to the participation of volunteers. Each volunteer is required to make a 24 month commitment to the project (Pecan Street Inc., 2012b). Pecan Street Inc. will uninstall all equipment after the testing period is over, meaning that volunteers retain no ownership over the technology. Volunteers are not able to choose what technology they can test; this decision is made by the research and development organization. Finally, data collected about a volunteer’s energy usage habits are jointly owned by the volunteer and Pecan Street Inc. (Pecan Street Inc., 2012b). While the members of the smart grid consortium receive some opinions and feedback from community members about the initiative, it is difficult to determine at this point, the extent to which citizens’ input has had an effect on the development of the technologies and how they will be used moving forward. While input from residents is sought, their ability to use and alter the provided technology in Illich’s sense of conviviality is hardly achieved. Control over the technology is not given to the residents, who simply provide data in a real-life demonstration, but have limited insights into the smart grid laboratory (cf. Collins & Pinch, 1998).

**Towards a Framework for Convivial Technology in Intelligent Urbanism**

In acknowledging the challenges faced by municipalities, urban planners, developers, designers and architects to understand how conviviality might be integrated into cities public and private spaces, we suggest and briefly describe four elements that are necessary to promote conviviality through technology. These factors can be used to design, deploy and maintain technologies, which empower...
residents. This will ensure their participation in the development and use of technologies, including a fair share of control and governance, and allow them to fully utilize technological resources to improve their quality of living.

**Facilitate communication, share information and enable access to data**- Communication and information sharing can take place at multiple levels and between multiple entities—between users, between users and systems, and between systems. Wirth (1938) addressed the way in which residents living in cities have an increased sense of isolation. Socially-oriented technologies can be leveraged to bring increasingly mobile, nomadic urbanites together both virtually and in the physical world. Further, data about and provided by the community should be made available to enable the building of convivial tools.

**Create choices and enable decision making**- Technology can help to provide more choices and alternatives for residents around transportation, healthcare, education, communication and more. Openness as an abstract concept allows residents to create and interact with technology, where possible, to reconfigure technology according to their needs. Given a plethora of options, technology can also collect, analyze and present insightful information to help individuals make sound decisions.

**Engage in commons-informed governance**- The participation of the people who will occupy the urban space and will be utilizing the technology throughout planning and deployment ensures that their needs and concerns are genuinely taken into account. Too often, tools and services are designed with a hypothetical user in mind, only to discover after the technology has been implemented that it is not the ideal fit for the particular context and intended users. The literature on common pool resources (CPRs) can provide guidance to inform governance regimes for convivial technologies in smart cities (Ostrom, 1990). Borrowing from CPRs, convivial technologies, to remain equitable, efficient and sustainable, must be governed by community members, who define and modify the rules around how the resource should be used (Hess & Ostrom, 2007). Similarly to CPRs, convivial technologies can be subject to social dilemmas; an appropriate governance regime will ensure that residents have a sense of responsibility towards sharing, control and the development of the technology that is embedded in the community itself.

**Enhance experience on an aesthetic, multi-sensory level**- Technology has created ever increasing mobility for its users and has enabled the manipulation of time and space in an urban environment. So often these benefits are considered through purely practical lenses, such as time and energy efficiency or business process improvement. But the experience of living in an urban environment for residents can be enhanced by technology on a more personal and emotional level. Technologies designed to encourage individuals to interact with their dynamic urban surroundings can create “an aesthetic experience rich with memory, imagination and brief encounters” (Williams, Robles & Dourish, 2008).

**Conclusion**

Given that society faces challenging problems in the 21st century that are reflected in the public as well as in the private spaces where individuals live and work, technology might be an ingredient in the solution of these multi-faceted issues. The same technology, however bears dystopian fears of surveillance and control. We chose Illich’s notion of conviviality to discuss and suggest a preliminary framework that should inspire one to think about how such technologies might be employed to enable a convivial life. However, we also have to take a step back and realize that technology alone is not the answer to that question, Illich invokes fundamental rethinking and reconfiguration of technology and beyond. In Illich’s terms, an intelligent urbanism that wants to achieve a convivial purpose must offer institutional choices to support a life of action and technologies that serve individuals themselves, rather than new ideologies and technologies that maintain a life style of consumption. Within new technology there is potential to enable a convivial life, that enables spontaneous, independent action while allowing individuals to interact and minimize isolation. However, technology such as such has an inherent potential in making individuals dependent, putting them in a position of pure consumption. Thus, the opportunity of the individual to choose, to be part of the decision making not only in living an urban life but also in creation and maintenance of urban life is essential. Conviviality in intelligent urbanism comes down to questions of governance. To create technologies that adhere to convivial living, governance of those technologies is crucial.
References

Pecan Street Inc. (2012a). What is Pecan Street Inc.? Retrieved from: http://www.pecanstreet.org/about/what-is-pecan-street-inc/
Information Management and Communication for Dementia: Preliminary Research from China

Claire L. Barco  
University of Michigan  
cbarco@umich.edu

Koji Yatani  
Microsoft Research Asia  
koji@microsoft.com

Yuanye Ma  
University of Michigan  
yuanyem@umich.edu

Candra K. Gill  
University of Michigan  
ckgill@umich.edu

Joyojeet Pal  
University of Michigan  
joyojeet@umich.edu

Abstract

In this note, we discuss technology use for professional dementia care in China. Caregiving for people with dementia is particularly stressful, and research has shown that technology could alleviate some of the stresses by facilitating the communication and information management around people with dementia. There are several problem points in the communication channels between the main stakeholders – seniors with dementia, professional caregivers, family members and the medical administration. We argue that research on this topic is particularly timely in developing countries like China because they face an aging general population and limited availability of caregivers and technologies. We present preliminary findings from a study on care facilities in Beijing.

Keywords: dementia, elder care, technology, China, developing countries

Introduction

Caring for people with dementia presents many challenges for the stakeholders involved. Elders with dementia may wander, get disoriented, upset with daily tasks, or in cases of advanced dementia have difficulties recalling their family members or even their own identities. These symptoms can cause stressful moments for the person with dementia as well as for their family members and loved ones. The affordances of communications technologies can play a role in both the management of symptoms and the flows of information related to dementia caregiving, particularly in places with aging populations and limited access to care. In this paper, we examine the arguments for this idea systematically and present some preliminary research results.

Our research, from January 2012 to August 2012, has focused on the use of communication technologies in the provision of care for dementia patients in China. China has an aging population and a rapidly expanding institutional senior care sector, with a transition from home-based care. Work in gerontology and information systems motivated us to investigate three areas where technologies may play an important role. First, we consider research on the emotional stress of caregivers and family members, relating to the time and communication-intensive nature of caregiving. Second, we look at research on information- and communications-systems related to caregiving, and third, research on media for people with dementia.

Professional and familial dementia care is often more stressful than other forms of caregiving for seniors – increasing risks related to employment complications, mental and physical health issues, and conflict (Ory, Hoffman III et al., 1999). For family members whose individuals with dementia live in professionally-run institutions for caregiving, the problem can be confounded by feelings of guilt (Martin, Gilbert et al., 2006) (Culter Riddick, Cohen-Mansfield et al., 1992).

Communication between stakeholders in particular has been recognized as a major issue in caregiving around dementia (Brodaty, Griffin et al., 1990). Studies have shown that structured information
gathering and maintenance can play an important role in better management of care, and there are moderate impacts particularly on caregivers (Powell, Chiu et al., 2008). Other studies have shown that the use of video conferencing and telecommunications systems helped professional caregivers obtain information from online forums and facilitated their communications with the family members of the individuals with dementia (Czaja and Rubert, 2002; Marziali and Donahue, 2006). Such systems also enabled interactions between caregivers (Bank, Argüelles et al., 2006), and potentially had impacts on the caregiver’s well-being (Eisdorfer, Czaja et al., 2003). Telecommunication systems have also been shown to positively impact caregiving burden among family caregivers (Finkel, Czaja et al., 2007). Finally, there is also evidence that computer-assistive cognitive interventions could play a role for the individuals with dementia to decelerate the advancement of their symptoms (Mahendra, Kim et al., 2005).

The third area of research that is relevant to our work is that of media and dementia. Work on the impacts of music therapy and familiar media has shown that music decreases aggressive behavior among people with dementia (Clark, Lipe et al., 1998; Koger, Chapin et al., 1999), and is an increasingly common practice in dementia care (Vink, Birks et al., 2003). Multimedia technologies for reminiscence therapy have shown mood improvements among individuals with dementia (Sarne-Fleischmann and Tractinsky, 2008), though the usability of these devices is often an issue. Work on designing interfaces usable by people with dementia has shown mixed results – while a number of cognitive prostheses (Alm, Astell et al., 2004) and interfaces (Boer, 2008; Wherton and Monk, 2008) have been tested for specific purposes, there is much variance on individuals’ ability to use these devices based on the progression of their dementia.

In summary, there are multiple levels at which technology interventions can be useful in both communication and the day to day experiences of well-being among individuals with dementia and their caregivers. However, little research has explored the feasibility of such technologies in the developing world where access to resources and media could be limited and is often different than in the West and relationships between caregivers and family members differently structured.

Stakeholder Groups in Dementia Care

There are four primary stakeholder groups involved in dementia care; persons with dementia, their family members, occupational caregivers (including nurses and personal caregivers), and the administration of care facilities. While there is need for communication between stakeholder groups, there is also need for communication within each stakeholder group. Figure 1 here illustrates the communication channels.

Communication Breakdowns between Stakeholder Groups

Many occupational caregivers (especially nurses) do not have the time to sit and chat with their patients on a regular basis. Nurses often do not know enough personal information about their patients in order to have meaningful communication even if they do have a free moment. Further, the increasing pressure on elder care can result in a lack of sufficient training for the employees of care facilities, which leads to impatience and a dearth of readily available communication strategies. Economic pressures on family members, particularly in countries like China, limit family members’ ability to spend in-person time
either with the seniors. Limited communicative engagement for seniors could increase the sense of loneliness, whereas for family members there is a risk of increased sense of guilt.

There is very little communication between occupational caregivers and family members. In general, there are no channels for communication built into the daily care schedule, and all communication about the elder is passed to the family members through the administration on a less than regular basis. This complicated method of communication leaves family members without regular updates on the person with dementia’s status, and is also a lost opportunity for family members to provide occupational caregivers with information about the patient’s personal history, likes and dislikes, and general personality. In practice, we find that occupational caregivers have to learn this information little by little from the patient.

Because occupational caregivers are focused on meeting the needs of the body, they do not track information about successful interactions with patients with dementia in their care. Information about what a patient likes to chat about is only kept in the minds of the caregivers, is not officially recorded, and there is no tracking to see if increased communication helps maintain a given patient's cognitive ability.

Communication Breakdowns within Stakeholder Groups

We find that there are no standards for occupational caregivers to share information about their patients between themselves. Basic information about health and medication status is shared through official channels such as end-of-shift report or health records, but information about the patient's personal history, personality, likes and dislikes are not shared in any official capacity. This is a lost opportunity to share helpful tips that can be recorded for new employees or shared with family members.

Family members often need to connect with others dealing with dementia's impact in the family, much in the same way that family members dealing with other conditions do. In-person support groups are rare, even in rural and low-income parts of the industrialized world, and this is a particular challenge in resource-constrained countries. Finally, individuals with dementia living in care facilities have social needs such as meeting other elders and people with shared interests, which can be difficult given their condition. Though professional caregivers and family members could serve as conduits for these connections, they are constrained by time and the lack of easy tools to do the same.

The Potential of Technology in Communication within Stakeholder Groups

Research has shown the value of forums for caregivers (Bank, Argüelles et al. 2006), and their use in dementia care can help not only in exchanging work-related information but also in creating social connections between individuals dealing with a similarly stressful work situation. Even in middle-income countries like China, the high mobile penetration supports such interactions on mobile platforms.

Similarly family members dealing with dementia can find community among others dealing with the same situation. Given the complex nature of direct information flows between the family and institutional care services, online forums can be targeted to needs and repurposed to communication or information, depending on the individual.

Finally for persons with dementia living in an institution, technologies can provide a platform from which patients can get to know each other. Residents could see if there were others in the home that had similar interests and send them messages or seek them out to build a relationship, share media, and organize social events. Systems for messaging within the home could be particularly useful for seniors with mobility issues. In order to combat feelings of low self-worth seniors with dementia can suffer, systems could be designed with small tasks that once completed could unlock points. These points could be redeemed for things like extra dessert at dinner or, if systems are developed in partnership with foundations, points could be donated towards a dementia research institution or other charitable cause of the senior's choosing.

Dementia Care in China - Research Overview

In preparation for our research in the summer 2012, our team looked into existing technologies for dementia care throughout the world. We found a number of reminder systems and GPS technologies already in existence, some already commercially available in China or provided to elders through the local government. For example, the Shanghai municipal government began in 2010 to provide and test GPS
enabled wristwatches for elders with dementia (Wang, 2010). Two missing pieces were systems that enabled people to maintain their social life — thus connections with people, music, and personal media — and systems that helped facilitate and integrate information between the various stakeholders. Using this as the point of departure in planning our research, our interviews focused on the communication chains, and if and how technology may play a role in making these smoother.

To deepen our understanding of communication between persons with dementia and occupational caregivers, we performed needs assessment on communication between the two groups. We conducted semi-structured interviews with 11 nurses (all female; P1 – P11) at three different elder care homes in Beijing in order to understand their perceptions regarding dementia and the need for daily, quality communication. Interviews were conducted in Chinese, and transcribed verbatim into English. Two of the authors then performed open-coding analysis.

Findings

We have found that our interview participants, while understanding the need for quality communication with their patients, face many difficulties establishing connections that would make this possible. As prior research pointed out, our participants do not have much time set aside in their daily schedule to sit and chat with the elders in their home. We also observed that seniors must be reassured or reminded of simple things on a regular basis. While some participants expressed frustration over this, others reported developing their own personal strategies for handling these situations, such as distracting the senior with a new topic. Other strategies included coaxing, lying, or ignoring the problem. Participants indicated that they did not always feel these strategies were ethically correct, but indicated that they were used out of necessity. As P1 pointed out:

“Just like coaxing a child, coaxing and cajoling I guess, sometimes I feel like it might not be right to lie to the seniors, but you have no choice, you cannot communicate with him in the normal way.” [P1]

In areas like China, institutional caregivers may not be knowledgeable of or trained in communication strategies for seniors with dementia, compounding the existent dearth of caregivers. Technologies designed for on-the-job training would be well suited to areas with a similar lack of knowledge and resources.

We also found that choosing topics carefully was very important in avoiding emotional stress for the seniors. Participants indicated that some elders enjoy talking about past successes, but others may not want to discuss the past, as this could bring up difficult or painful memories for them. This was apparent through the contrasting comments from P4 and P1.

“Try to speak to her the way she likes, try to ask her ‘how was your experiences in the chorus? How was your cello playing experience?’ those are topics she’d enjoy.” [P4]

“It would not be appropriate for us to ask about their past stories either, because in my own understanding, possibly some of the past experiences would be painful which they might not want to recall, at this stage of life they might not necessarily want to think about those histories again.” [P1]

Appropriate topics were usually discovered through trial-and-error, as there was limited communication between family members and the nurses. Family members did on occasion provide helpful information to the nurses while visiting the home. For example, when P10 was asked about how they knew a specific elder was once a teacher, they replied:

“I know from his son when he was visiting.” [P10]

Additionally, many of our participants reported that their elders do not have mobile phones. This is due to the fact that persons with dementia might not remember that they recently spoke with their family, and would call them many times throughout the day. Any communication technology designed for use between persons with dementia and family members should be carefully designed as to reduce undue burden and stress on the family members as well as persons with dementia.

Conclusions

Our results suggest that enhancing communications is likely to smoothen some painful areas particularly relating to contextual reminders, family connections, and conversation prompts. Although our starting point is examining the possibility of technology use, we are not constrained to this. While our
initial design ideas were around the possibility of dementia care for middle-income regions, the extensive use of technology in care facilities for seniors in developed countries opens us to the possibility of similar work in the States. At present day prices, a basic tablet device which can be used to support many such communications can cost as little as a few hours of professional care while a mobile device would be even cheaper.

iSchools have as part of their research agenda areas such as technology adoption, contextual inquiry, information management, and health informatics. The communications needs of dementia caregiving draws from work in each of these. Using our results, in the coming months we plan to begin preliminary designs for use in the institutional setting.

References


Exploring Hybrid-Economic Communities and the Technology-Mediated Identities Performed There

Warren Allen
Drexel University
Philadelphia, PA
wsa25@drexel.edu

Abstract
This research note describes an ongoing ethnographic study of a professional, knowledge-sharing community of information technology professionals. The overarching goal of the research is to understand the use and consequences of social information systems designed for recognizing and rewarding contributions to professional communities within and beyond the workplace. The research presented in this note explores the following question: How does technology mediate the performance of professional identity, and what role is played by the social institutions – e.g., the firm as employer, the firm as a market actor, the professional community, etc. – in enabling and constraining the construction of a professional identity?

Keywords: social media, self and identity, professional communities, communities of practice

Introduction
The ongoing research presented in this note research explores the role of technology, social institutions, and individual agency in the construction of professional identity, and aims to understand the use and consequences of social information systems designed for recognizing and rewarding contributions to professional communities within and beyond the workplace. The guiding research question is as follows: How does technology mediate the performance of professional identity, and what role is played by the social institutions – e.g., the firm as employer, the firm as a market actor, the professional community, etc. – in enabling and constraining the construction of a professional identity?

Mediated performance of professional identity
Increasingly, everyday performance of the self (Goffman, 1959) occurs in technology-mediated social contexts where the capacity for a mediated audience to develop impressions of an actor exists to the same degree as in face-to-face interactions (Walther, 1996), and the motivation and opportunity for deception and impression management through selective self-presentation remains (1996). Online self-performances are afforded opportunities for information-based “social grooming” (Donath, 2007; Tufekci, 2008) by the design of the mediating technologies, and leave behind “a multitude of data traces [which] do not merely document our passage in life’s play but mediate our parts” (Hogan, 2010). Existing research on identity in professional contexts has explored its role in establishing employee commitment and identification with the firm (Knippenberg & Sleebos, 2006; Vough, 2012), and particular forms of identity work in professions such as Episcopal priests (Kreiner, et al., 2006) and cruise personnel (Tracy, 2000). In the context of professional work, Ibarra (1999) theorizes that junior professionals develop a professional identity by observing role models of potential identities, experimenting with “provisional selves,” and evaluating their experiments against internal standards and external feedback. Hara (2007) examines

Copyright is held by the author.
the role technology plays in supporting work practices and identity formation of communities of attorneys, finding that while technology fostered instrumental, communicative, and discursive aspects of professional practice, it was not found to foster community or sharing of cultural knowledge. Dabbish et al. (2012) describe "social inferences [...] such as inferring someone’s technical goals based on actions on code" made by open-source software developers from activity information mediated through GitHub, a Web-based open-source code repository. These inferences informed strategies for advancing technical skills and managing reputation.

Hara (2007) found that “[while] younger attorneys learn how to be public defenders by observing more experienced attorneys and by talking with them,” the use of information technology within those communities did not facilitate the transfer of cultural, tacit forms of knowledge. However, Dabbish et al. found that software developers used features of the GitHub website to learn better ways to code and to gain access to superior knowledge by watching how others code, what they paid attention to, and how they solved problems (Dabbish et al., 2012). In other words, in some contexts (e.g., GitHub developers) the technological mediation of social signals was found to contribute to the transfer of social knowledge and learning of cultural, tacit forms of knowledge; in other contexts (e.g., Hara’s attorneys) the use of information technology may fail to play and effective mediating role. Aspects of the sociotechnical situation will affect outcomes, including the technology’s design, the types of social information it can mediate, and any number of social contexts that shape the technology’s use.

Research described in this note shares with prior studies an interest in professional identity, its immediate and mediated forms, and the institutional forces that enable and constrained it. Presented in this note are preliminary findings from ongoing ethnographic work, with a focus on the social information mediated by certain types of information systems that are designed or used to explicitly give – or implicitly give-off – social information relevant to the self and the construction of a professional identity.

**Methodology**

The findings discussed in this note are derived from analysis of field notes collected during participant observation conducted at events related to CollabTech WorkSphere – a popular enterprise-level software system for information management and collaboration – and in online, social media settings between July 2011 and February 2012.

WorkSphere is a closed, commercially developed, proprietary software product surrounded by a knowledge-based economy of consultants, system integrators, developers, designers, and others. Community members share knowledge through participation in a variety of face-to-face national and regional conferences, local user group events, and in online forums and social media spaces. Many of the community’s more active constituents (both individuals and organizations) actively discuss WorkSphere and other topics on Twitter and, to a lesser degree, on the Yammer platform.

Knowledge-sharing through community participation (online and in face-to-face settings) occurs on a voluntary basis and as a consequence of employment at CollabTech or one of many software, services, and consulting firms operating in the WorkSphere market. Many of the observed knowledge-sharing and learning opportunities doubled as marketing presentations; not for CollabTech itself, but for companies that operate in the WorkSphere market. Additionally, though WorkSphere is itself a closed-source, commercial software product, many software tools are developed by volunteer developers and released under open-source licenses through a free-of-cost, open source project hosting site supported by CollabTech. The interaction of the market and the community remains a focus of the ongoing ethnography, and is a component of the following research vignette.

**Ethnographic Data**

This section contains a research vignette derived from fieldwork data. Following the vignette is an analysis of the role of the CollabTech VIP award in the performance and mediation of identity, as well as the dynamics that are observed between the market, the commons, and the award itself.

**Worksphere VIPs: The Passion of the Community, or the Heartbeat of the Market?**

CollabTech describes the goal of its “Very Important Professional” ("VIP") as the recognition of exceptional community leaders from around the world who voluntarily share their deep, real-world
knowledge about CollabTech technologies with others. The nomination process— which can be initiated by anyone in the WorkSphere community— asks for basic information about the nominee and specific examples of online or offline community activities in which the nominee participates. Beyond CollabTech’s statement that “to receive the VIP Award, you must be passionate about CollabTech technology,” comments made by individuals in the community suggest, “it’s not clear what criteria CollabTech uses to choose who gets the award.” [anonymous 27:1]

Recently, a software developer (“Bob”) had his VIP award renewal denied, launching a discussion on Twitter and in multiple blogs about experiences with the VIP program. As the sole developer of a popular open-source add-on to CollabTech’s closed-source programming language, Bob’s values about transparency—and the potential for conflict with the VIP program—are evident in his critique:

One of the problems with the VIP program is that the whole thing is basically a mystery. Here’s where I first knock heads with the program. I value transparency and openness, even if it’s difficult or sometimes painful. The VIP program does not value openness. That’s why it’s basically a mystery how you get nominated for a VIP or what you have to do to get one. [Bob, P3:18]

Others have suggested that there are certain ways to become a VIP even though there are no official criteria. Across multiple blogs and blog comments regarding the matter, a range of views reflect diverging thoughts and opinions about the program:

VIP and the Certifications are about promoting CollabTech products. That’s it, nothing else really. The reward I get for doing the Certifications isn’t the cert, it’s the fact that my company will pay me a bonus. Why do they care? Because they want employees with certs to maintain their CollabTech status. Why do they want that? Because of sales. Makes sense, and a nice cooperative relationship. I win, they win, CollabTech wins. [Sam 3:107]

Many years ago the VIP award was for answering questions on Compuserve and then NNTP newsgroups. […] If you had a website/blog or http forum, good for you, but you wouldn’t get an award. It was about answering the direct posts of people who needed and answer. So things change, but we don’t have to like it. Today any douchebag with a blog could get one. Marketing is king. [anonymous 13:6]

Not all bloggers and commenters view the program negatively or with such skepticism:

Don’t forget that being VIP is a PRIVILEGE, NOT A RIGHT. Sure, you built a cool framework, but in all fairness it was probably your bitter attitude that you’ve expressed here as well that made your VIP not be extended. [Fred 3:58]

Achieving a VIP award means you did *something* to deserve it. You were involved with user groups, code camps, speaking engagements, posting on forums, writing blogs, writing books, etc. The fact that CollabTech is recognizing your accomplishments isn’t something to take for granted, regardless of how secretive and closed the selection process is (a favourite complaint of many). [Donna 16:1]

At a point in the VIP nomination process, each nominee completes a spreadsheet or submits information by way of an online form that has been the target of criticism:

When I came up for renewal, I had to *defend* why I should have my award renewed using a bad Excel spreadsheet and a really bad online form which don’t even allow me to capture my real contributions. [Bob 3:9]

Don’t get me started about that “silly” Excel form… I complained about that one, refused to fill it out and sent a bulleted email instead. [Marco 3:30]

Though not all comments observed during the study were negative towards the spreadsheet or its role in recognizing community contributors:
IMHO, it's reasonable for [CollabTech] to ask what you did to be a VIP once a year - it's hardly a monumental task (although the site/spreadsheet is ridiculously painful to work with). [Reggie 3:94]

I discussed the nomination process with Blake, a recently-awarded WorkSphere VIP. In addition to sharing information about the selection process, he included the document he sent to the VIP Program representatives instead of the spreadsheet. His 10-page document outlined his many activities in the WorkSphere community throughout the year. The document bears comparison to the following criticism from Bob E.'s blog post mentioned above:

_The spreadsheet you have to fill in already shows what kind of people CollabTech really wants to be a VIP: people who volunteer to do their PR, their evangelisation, the marketing of their products, in short: sales people who don't cost a dime! (well, perhaps the MSDN licenses, but they actually don't cost CT any money) So for CT, a person who did a lot to spread the CollabTech gospel among fellow community members, that kind of person becomes a VIP. Not a person who actually did something for the community, like contribute a lot. [Bob 3:22]

Bob brings into contrast "a person who actually did something for the community, like contribute a lot" with "a person who did a lot to spread the CollabTech gospel among fellow community members." (To wit, Blake’s professional title is Evangelist, not an uncommon term and job title in the WorkSphere community.) Bob’s scheme of perceived value is emblematic of his role as an open source software developer; i.e., he values code contributions above all other forms of contribution to the community.

The content of Blake’s self-designed nomination form reflects drastically different values: He also produced a series of short videos highlighting community experts sharing their answers to the question, "what is the one thing everyone needs to know about WorkSphere 2012?" and community-spoofing videos such as “Behind the Music: The biggest fake WorkSphere band, ever.” Blake has been the most prolific speaker at user groups and worldwide community events in the last year and he is often seen at events handing out t-shirts capturing a range of community memes and inside jokes. Finally, he authored a “free” e-book (it “costs” the reader an email address, paid to Blake’s employer) the title of which — _Inside the WorkSphere Community: 4 Strategies for Building Your Personal Brand — reflects his perspective on the value of doing pro-active forms identity work he (among others) refers to as personal branding.

Ultimately, both Bob and Blake received a 2012 VIP Award; the former reluctantly, the latter with conspicuous fanfare from his employer, Excelsior, a software and services firm that released a press release with the following headline:

_Excelsior Inc. Evangelist Blake Doe Receives CollabTech’s VIP Award: Doe Recognized as Top-Tier Technology Influencer and Community Leader in the WorkSphere Category [Excelsior, 2:1].

The firm’s press release continues, claiming that the award is “also an acknowledgement from CollabTech of Excelsior’s broad contributions to the WorkSphere community” [Excelsior, 2:2]. Having attended multiple of Blake’s sessions and speaking with him about the award, there can be as little doubt about his personal loyalty to the community as there is about his company’s desire to capitalize on it.

**Discussion and Conclusion**

Amitai Etzioni writes that “in any relationship or community [there is] some vague sense of appropriate reciprocity, of the need to contribute to a climate of mutuality. But basically people help one another and sustain the spirit of community because they sense it is the right thing to do” (Etzioni, 1994). Findings suggest there is indeed a “climate of mutuality” in the WorkSphere community, but it may not be the dominant climate. Non-market forces like volunteerism, knowledge-sharing, and social reciprocity are evident throughout the WorkSphere community, and shape the structure of the community. However, the logic of the marketplace is also observed throughout the community at events and in online forms of participation. The VIP award is a product of the resulting “hybrid economy” (Lessig, 2008). Despite CollabTech’s portrait of a VIP as an altruistic contributor independently sharing his or her expertise within the community with no motivation but the passion for WorkSphere, the award is equally effective in its capacity to mediate the ostensive market capabilities of the firm.
Beyond Community and Social Capital

To further the understanding of the factors that shape self-performance and identity-building behavior in a hybrid-economic community the conceptual framework for analysis will be expanded in two ways. First, the concept of social capital – a popular one in the study of social media use by young adults (Ellison, et al., 2007; Steinfield, et al., 2008) and in technology-enabled knowledge exchange (Inkpen & Tsang, 2005; Wasko & Faraj, 2005) – may be inadequate for a social context in which multiple forms of capital co-operate and interact. Therefore, continued analysis will adopt a conceptual framework derived from Bourdieu’s work on the conversion of economic, cultural, and social forms of capital (Bourdieu, 2008). An example of a conversion of capital is found in the appropriation of Blake’s social and cultural capital by his employer, Excelsior, Inc. By using forms of economic capital (namely, the money used to employ Blake and otherwise financially support his participation in community events) Excelsior benefits from the social and cultural capital built up over time that is, ostensibly, the product of Blake’s community participation. Excelsior cannot realize the benefits of a VIP award through a single form of capital (the company cannot buy the award; not directly, at least) and as we have seen, the award is coveted by many in the community (and dismissed by others) not because of its capacity to recognize social capital but for its capacity to be converted into economic capital; i.e., there are few underpaid VIPs if there are any at all.

Secondly, a sociological lens will be adopted, constituted by three related concepts: Gemeinschaft (associated with common ways of life and beliefs, strong ties with fewer individuals, emotional bonds, and continuity [Brint, 2001]), Gesellschaft (associated with dissimilar ways of life and beliefs, weak ties with many individuals, and temporary and regulated arrangements [Tönnies, 1955]); and Pseudo-gemeinschaft, “the feigning of personal concern with the other fellow in order to manipulate him the better; [i.e.,] the mere pretense of common values in order to further private interests” (Merton, 1975). Gemeinschaft and Gesellschaft are not dichotomous concepts belonging strictly to their respective forms of social organization; rather, they are dynamic attributes of the human experience that “are found interwoven in all kinds of associations” (Tönnies, 1955). I expect that Pseudo-gemeinschaft is also woven into many transactions that occur within the WorkSphere community. Furthermore, I expect that these concepts will aid the development of a framework for exposing the dynamics of community participation, knowledge-sharing, and the conversion of capital in the WorkSphere community without forcing all participatory behaviors into either a market-driven economic exchange or a community-based sharing exchange.

Finally, social media technologies continue to infiltrate the workplace as the latest wave of enterprise-level communication, collaboration, and knowledge management support systems. Among the most recent are social information systems (Allen, 2010) designed to measure – and often incent and reward – such social constructs as “influence” and “reputation” based on social media and online community participation. Such systems emerged recently on the public Web, e.g., Klout, PeerIndex, and the many “badge systems” (Antin & Churchill, 2011; Halavais, 2012) designed into Web-based community platforms like StackOverflow. These systems have also entered education and training areas where badge systems and platforms promise new but untested modes of engagement (Muntean, 2011) and assessment (Lee & Hammer, 2011; Simões et al., 2012). These systems are making their way into workplace-based social computing (e.g., Maybury, 2002) ushered in under the idiom of Gamification, “the use of game design elements in non-game contexts” (Deterding et al., 2011). How these systems are designed and the information they are designed to mediate will shape (and be shaped by) the ways in which users appropriate the technology to perform their identities.

Preliminary analysis of the CollabTech’s VIP award – itself a system ostensibly designed to incent and reward knowledge-sharing and community participation – suggests that the diversity of motivations in the WorkSphere community is reflected in the varied attitudes towards the VIP program. By expanding the concept of capital exchange to include conversion of various forms of capital and by re-framing the analysis of community participation to view the WorkSphere community as a hybrid-economic structure, a more nuanced view of the forces that enable and constrain identity performance is possible.
References


It Takes a Jerk to Make a Conversation into an Archive

Julia Bullard
University of Texas at Austin
julia.a.bullard@gmail.com

Abstract

This preliminary study examines an online community’s open censorship moderation practice, revealing a concern for the quality of the archive produced by discussion board discourse. Moderator comments in a censorship sub-forum are examined to determine the rationale behind otherwise vague published rules. Instead of motivating participation with supportive administration strategies, the strict moderation practices in the elitistjerks.com (EJ) forums intervene in conversations in ways that prioritize the intelligibility and accessibility of the archive by being a “jerk” to bad contributors.

Keywords: online communities, moderation, censorship, archives

Introduction

The relationship between an archive and a discourse is fundamental to understanding infrastructures of knowledge. The relationship is mutually constitutive: an archive supports a discourse, but, over time, a discourse produces contributions that are added to the archive. This process is highly visible in online discussion boards in which all conversation occurs through shared records (Millen, 2000). However, not all engaging and valuable conversations create intelligible and accessible archives; rather, some intervention is needed to produce a usable archive out of a community’s discourse (Hansen, 2009). I examine an online discussion community, highlighting the centrality of the quality of the archive left by ongoing discourse to shaping the practices of moderation. Interestingly this exposes a clear value in elitist, seemingly high-handed behavior that runs counter to the idea that moderators in online discussion forums are primarily concerned with motivating participation, a common assumption in the online communities literature (Kraut & Resnick, 2012).

The elitistjerks.com (EJ) discussion board is devoted to the practice of theorycrafting, in which video game players engage in rigorous experimentation and modeling of the game software’s underlying rules to support optimal in-game practices. Theorycrafting occurs predominately through discussion in third-party player forums in which players set the rules by which information is shared and organized. This paper looks at one of the major modes of control used in these sites, that of human moderation of user posts. The rules that govern submissions in this community are explicitly non-negotiable, and there are no traces of debate to examine regarding their meaning (Butler, Joyce, & Pike, 2008), but moderator comments on infractions reveal their own interpretations of otherwise terse policies. I will examine how the published infractions in EJ invoke a more substantial code of conduct than the published rules suggest, and indicate a concern for the maintenance of an intelligible and accessible archive.

The EJ moderation practices comprise a knowledge management strategy that reflects the information needs of its members. I propose that the EJ forums be regarded as a successful example of user-driven management of information, one that suggests that user-generated content can be organized effectively absent professional intervention. User-driven information management strategies are not limited to folksonomies or tagging; the moderation practices of the EJ forums demonstrate a different mode of information management in information creation.

Acknowledgement: Thanks to Dr. Lynn Westbrook, Dr. James Howison, and Dr. Melanie Feinberg for their help with multiple versions of this paper.
Bullard, J. (2013). It takes a jerk to make a conversation into an archive. iConference 2013 Proceedings (pp. 583-588).
doi:10.9776/13284
Copyright is held by the author.
Related Research

While there is little research that directly addresses the effect of moderation practices on discussion boards resembling EJ, three growing bodies of knowledge suggest ways of understanding the phenomenon: e-democracy, information creation, and theorycrafting.

E-democracy and Open Censorship

The impact of moderation practices on the quality of online discussion is particularly important to the potential for e-democracy, the role that internet technologies might have in supporting information access and open communication necessary for civic engagement. E-democracy research has shown that moderation practices are a delicate but vital component to a well-designed deliberation space (Wright & Street, 2007, Wright, 2009, Coleman & Moss, 2012). Scott Wright (2009) warns that silent or “covert” moderation in particular can “create a conspiratorial atmosphere as messages are removed without explanation” (p. 237). “Open censorship” is seen as a constructive alternative in which a moderator replaces posts that have violated published rules with messages explaining the nature of the infraction (p. 236). Open censorship manages the conversation by pruning inappropriate content and creating concrete examples of published rules.

Information Management for Intelligibility and Accessibility

Moderation can also be a preemptive means of information organization for retrieval. Huvila (2011) states that the challenge of managing a wealth of user-generated information must be addressed by improving information creation and information organization. A focus on information creation can facilitate information retrieval as users-as-authors learn how to “reduce the complexity of information and its expressions” (p. 238).

Huvila (2011) emphasizes that problems of information retrieval have their roots in information creation. “If the motivation for creating information were maximum intelligibility and accessibility”, he states, “problems would hardly exist to such a degree now” (p. 240). Instead, other motivations determine the form of documents, such as the need to manage daily life, to communicate status, or to signify an accomplishment (Trace, 2007). The more narrowly a discussion forum defines its purpose as intelligibility and accessibility of information, the more of these alternative motivations will be weeded out in their moderation practices.

Theorycrafting

Theorycrafting, as the pursuit of accurate models of hidden game rules, values intelligibility and accessibility above the other information creation purposes posed by Huvila and Trace. Personal motivations are subordinate to the community goal of producing reliable and actionable knowledge. Among the information resources available to World of Warcraft players, the EJ forums’ are regarded as the most reliable, which has been attributed to their strict posting standards (Thomas, 2009, p. 40). The major theorycrafting threads of EJ, which are grouped according to in-game class, contain approximately 750 threads totally nearly 400,000 posts.

The primary users of EJ and theorycrafting data are hardcore players who engage in instrumental play, defined as the pursuit of the optimal mode of play directed at the most difficult goals in the game world (Taylor, 2006). As an example of “productive play”, theorycrafting resembles the work of engineers experimenting on tools and tuning them for greater performance (Nardi, Pearce, & Ellis, 2008), or of scientists engaging in scholarly communication (Nardi et al., 2008, Steinkuehler & Chmiel, 2006). The collaborative and rigorous nature of these tasks requires different modes of communication than usually found in discussion forums, including greater intelligibility and accessibility.
Current Study

This study will examine how the practice of open censorship transforms the discussion board into a suitable communication system for theorycrafting, and how moderator comments on infractions both educate users on proper conduct and show concern for the quality of the archive.

A close reading of open censorship traces – found in the “Banhammer” sub-forum – will answer the following questions:

• How is open censorship used to create and maintain a usable archive out of conversation?
• How can values such as information accessibility, intelligibility, and critical thinking be expressed through open censorship?

Research Design

As a pilot study to a more comprehensive analysis of the Banhammer sub-forum, I selected a 3 week section of Banhammer threads. On the EJ forum index, the Banhammer sub-forum is explained as:

“There are a) those who A. can’t contribute; or B. have broken shift keys; or C. think, of course, noone, or lol are actual words, will eventually find themselves in here.”

Each Banhammer thread represents one moderator’s response to a single user’s post, and includes a description and assigned penalty for one or more infractions against the site’s rules. A “Forum Rules” page linked from each sub-forum gives short explanations of each rule (Forum Rules, 2010). The Banhammer subforum contains approximately 50,000 threads, the vast majority of which consisting of single posts. The three-week period between March 21, 2012 to April 11, 2012 included 69 threads.

Thread contents were first coded according to the rules broken and the severity of punishment. While the sample included infractions for nine out of the ten forum rules, moderator comments were only coded for infractions of the vaguest rules:

6. Do not post unless you have something new and worthwhile to say. [Not Worthwhile]
7. Do not beg for hand-holding. [Begging]

Moderator comments were coded in a grounded theory approach, in which labels were suggested by the text rather than matched to an external taxonomy (Glaser & Strauss, 1967). At this pilot stage of the analysis, the focus was on deriving broad concepts from the moderator’s explanations and proposing a tentative thesis as to the larger purpose of this open censorship strategy.

Analysis

Clutter

One of the moderators posting “Not Worthwhile” violations used the term “clutter” to describe the problem with offending posts, while others described similar offending posts as “useless”, “adding very little”, or content they “don’t care” about. These descriptions occurred in approximately one quarter of the threads. The idea of clutter is particularly relevant to the function of the discussion board as an intelligible, readable, accessible body of knowledge. Some sub-sections of the discussion board include a “sticky” post advising contributors to a topic thread to “read the last 5 pages to be sure the topic you had in mind hasn’t recently been covered”. Clutter posts dilute the informative content of topic threads so that it is less likely that “the last 5 pages” will include recent answers to good questions, and make the experience of reading through a thread in preparation to ask a question a more tedious process, which in turn might lead fewer posters to follow this directive, which would lead to further clutter posts, ad nauseum. Clutter posts endanger the community standard that a finite amount of reading is a worthwhile investment and will answer most questions. Clutter posts can also be seen as a social dilemma phenomenon: an EJ user with a specific question can put in a minimal amount of effort and get an answer to their very specific question, or they can read through the accumulated information and deduce from general principles a specific solution, and avoid reducing the precision ratio for the next individual who conducts a search on a similar issue.
Effort & Questions: Search and Research

Half of moderator comments accuse the poster of not putting in enough effort to find relevant information before posing a question to the community. Often, this is described in terms of “doing your research”, “searching for the answer”, or “read more”. Moderators also chastise posters for not using appropriate external resources, including database websites and tools created by community members to perform calculations based on player data.

Moderators also single out questions that – while they may represent an information need appropriate for the thread – are not constructed with enough effort to be answerable. Usually this is described as “not enough” information in the question. Moderator comments describe a phenomenon similar to Taylor’s (1962, 1968) compromised information needs. The posters have access to more contextual information about their information need than they share, producing questions so general that they are either unanswerable or have the appearance of “Begging” because there is no evidence that the poster has yet put any effort into figuring out the answer for themselves. Unlike the traditional in-person reference interview, the discussion board post is not synchronous, takes longer to negotiate back to a formalized need, and takes place in public document where such an exchange would create clutter. By removing these posts, moderators pre-empt any exchanges to clarify the question, and dissuade users from asking for help before investigating on their own.

Effort & Answers: Hard Data

Effort is also an issue in answers to questions. A smaller cluster of “Not Worthwhile” Banhammer threads deal with answers that do not meet a standard of effort and quality. A representative example of this type states “Test before posting, not the other way around”. Moderator comments repeatedly call out posters who engage in speculation, who present anecdotal evidence, or who rely on gut feelings as support for their conclusions. These posts are classified as “not adding anything new or worthwhile” because they do not meet the standard of reliable information valued by the community. “Bad” answers, at best, will be part of the clutter that increases the tedium and effort required to read through a thread or reduces the precision of a search on a particular topic. At worst, they are not differentiated from conclusions based on testing and “hard data” and lead players to make bad decisions in-game. EJ, as a discussion board, does not have the “up/down” voting mechanics or reputation scores of Question & Answer sites. There are few cues about authors that signal cognitive authority, and users must judge posts on their own merit, largely relying on their presence in an EJ topic thread as a sufficient indicator of quality.

Punishment & Pruning

The difference between the rationale behind the two kinds of infractions can help explain the relationship between the rules and the severity of the punishment given. Half of the “Not Worthwhile” posts were Warnings with a 0 Points value, whereas almost all of the “Begging” infractions received a 1. Many of the “Not Worthwhile” Banhammer threads can be classified as clutter. They do represent bad habits of posters, but largely well-meaning habits – expressing appreciation for useful information or mistaking the proper thread for a particular kind of question. Those receiving 0 Points describe easy-to-fix issues, largely matters of understanding policy and etiquette. These posts are moved to the Banhammer to demonstrate the nuances of posting etiquette and to prune unnecessary posts from the topic threads.

The “Not Worthwhile” threads that do receive Infractions and a majority of “Begging” threads are more indicative of attitude and effort than etiquette. The moderator’s role is not only to remove the clutter from the flow of the original thread but to punish the poster for engaging in this type of behavior. These posts consist of “lazy” questions that ask the community to answer hyper-specific needs that are unlikely to be relevant to future users. They also include answers based on unscientific forms of evidence, such as speculation, anecdotes, and affect. This latter class of posts endanger the community by increasing the possibility that users will follow bad advice, reducing the overall reliability and reputation of EJ as an information resource.
Conclusion

This exploratory study has shown how the open censorship policy of EJ moderators has adapted a relatively unstructured discussion board system into a rigorous space for information dissemination that produces a searchable, readable, and usable archive for those users willing to put in a reasonable amount of effort. By strictly enforcing a rule against asking situation-specific questions, the moderators improve the precision of search functions within their site, and deter users who would avoid putting in the effort of conducting searches themselves.

It is telling that, despite the disadvantages, discussion boards remain a major location for theorycrafting information and activity. The collaborative writing and editing features of wikis are an obvious alternative, but an early attempt to collocate theorycrafting conclusions in a wiki-driven system was eventually abandoned in favor of maintaining the discussion-based format alone (Theorycrafting Think Tank, n.d.). A possible explanation is that forums support a conversational mode for scientific argument that is not featured in wiki systems. Theorycrafting continues to occur on a discussion board system that is familiar to the community, is highly accessible, and supports unstructured, informal contribution, but through human content moderation practices maintains the intelligibility and accessibility the community needs.

As a pilot study, this analysis was conducted on a narrow range of posts, only 69 out of over 50,000 in the sub-forum’s archive. These findings have informed an ongoing study, in which content analysis from multiple coders is being applied to a larger sample taken from a full cycle of activity in the site – between two major expansions of World of Warcraft. This larger study will be able to confirm the general concepts and priorities identified in this study, and to ask further questions. The larger sample will make possible analysis at the level of individual moderators, and whether interpretations, priorities, and policies have changed over the last 3 years of the site. It may also reveal how effective these policies have been in mentoring active contributors, or whether they merely chase off unwanted members. We must also investigate what about this community makes the elitist jerk management style work, while the literature suggests positive reinforcement and motivation.

At this stage, EJ can be understood as a successful and highly visible example of how strict moderation can adapt a seemingly ill-suited system to the particular needs of a community. For this “hardcore” leisure community, high quality, accessible information is necessary to support a high standard of play. It also supports the future “work” of theorycrafting in that moderation practices seem to support those who put in the most work – both by punishing contributors who fail to meet their standard of rigor and effort, and by disallowing questions that would take advantage of their generosity by subjecting them to ever-more-specific queries.

References


iPad as PIM Device: Input, Interrelation and Indexes

Paris Buttfield-Addison  paddison@utas.edu.au
Christopher Lueg  christopher.lueg@utas.edu.au
Leonie Ellis  leonie.ellis@utas.edu.au
School of Computing & Information Systems
University of Tasmania

Abstract

This paper presents preliminary results from components of a larger research study which explores the processes and challenges of information workers conducting personal information management (PIM) within offices. Here, we focus on the findings specific to users of iPad-style tablet computers with the Evernote PIM software; data reported was collected using several online questionnaires and semi-structured interviews; content analysis and open coding techniques were used to analyse the data. Our results suggest three themes — input, linking and tagging fatigue — that regularly impact users of this software, these themes are not specific to Evernote, and are common to many PIM tools. Our results give insights into the use of software for PIM, generally, and specifically the use of software on iPad-style tablet computers for PIM, and we provide suggestions for future work to improve tools and practices.

Keywords: personal information management, human-computer interaction, mobility

Introduction

Personal information management (PIM) is the process of acquiring, organising, maintaining and retrieving the multitudes of information items that an individual uses to fulfill their various roles, both professional and personal. Efficient PIM improves quality of life for individuals, through better use of time, energy and resources (Jones, 2007; Whittaker, 2011). The rise of consumer mobile computing has led to numerous tools designed specifically to raise the efficiency of PIM. Here we present preliminary results from a larger ongoing study of personal information habits by knowledge workers; this paper specifically discusses the methods by which users of Apple iPad tablet computers input information into PIM software on these devices, and how the choice of method affects the various stages of PIM. The PIM software tool primarily discussed here is Evernote, a combined service and cross-platform application geared around notetaking and archiving; Evernote supports folders, tagging, text formatting, images, photos, searching and archiving in notebooks (Evernote Corporation, 2012).

PIM is frequently examined through the lens of a variety of frameworks — this work applies the framework conceived by Barreau (1995), which encompasses the four stages of acquisition, organisation, maintenance and retrieval. The goal of the preliminary work presented in this paper is to provide initial analysis of results around the design of PIM software for the Apple iPad. Several themes relating the use of tablets for PIM are identified and introduced in this paper: data input, data linking, and tagging fatigue.

Methodology

The work reported here is derived from data collected from twenty-six face-to-face semi-structured interviews and sixty online questionnaire responses, both protocols taking place in 2012. Interviews were up to 90 minutes in length and revolved around the challenges office-based knowledge workers face in organising their personal information across physical and digital mediums; interview
participants are referred to as P1 to P26 in the remainder of this paper. The design of the questionnaire and semi-structured interviews was strongly influenced by past PIM studies (Malone, 1983; Whittaker & Hirschberg, 2001). Participants were aged between 21 and 68 years of age, and none were engaged in technology-related professions.

During the interviews, participants discussed how, when and why they used their iPad for PIM — the interviewer sought self-reported data from participants around activities in each of the stages of the Barreau (1995) PIM framework. An open coding-based method was used on the interview data, and the initial analysis here concentrates on the themes of input, linking and tagging fatigue, with an emphasis on deriving suggestions for design of future iPad, and other platform, PIM software. The coding scheme resulted in key themes, such as strategies, problems, tools, and challenges, which were used to mark up the data for further passes. Data from an online questionnaire is also reported, the online questionnaire related to the use of iPad-style tablet computers for PIM — results here report on the initial sixty participants in this questionnaire.

Findings

Our participants used their iPads for a wide range of PIM purposes, from creating long documents via the on-screen keyboard, to audio recording meetings, to drawing diagrams on the screens. All participants from the semi-structured interviews reported themselves as being active and highly regular users of the Evernote for iPad software. Here we present the three main themes developed from the initial analysis of the interview and online questionnaire data, those themes being: data input, data linking, and tagging fatigue.

![Figure 1. A sample of one stage of the coding process used in data analysis.](image)

![Figure 2. Themes identified, as related to relevant components of Barreau’s (1995) PIM framework.](image)
The first theme, data input, was identified as a significant point of consternation for users of iPad-based PIM software. The iPad alone effectively offers three means for inputting text: the on-screen keyboard, the cameras, and the inbuilt voice recognition of certain versions of the iPad. The revelation that the majority of our interview participants used the cameras as their primary input device for PIM using Evernote was intriguing; P6 provided a description typical of these participants: “the way I take most of my notes in Evernote is by taking pics [sic] of the pages in my notebook. It’s kind of painful to take a picture, save it, then take another one when really they’re all pictures of the same set of notes and should be stored as one note in Evernote”. The fact the our participants all seemed to rely on the camera as a primary form of data input — a primary part of the acquisition stage of PIM — is perhaps unique to the newest generation of tablet computers (i.e. iPad-style). While this data input technique could, perhaps, be equated to the use of a desktop scanner with a traditional personal computer, user interfaces and workflows within software, such as the Evernote app discussed here, are not at all geared towards this use-case.

The second theme identified was that of linking individual records together; as noted above, partici- pants frequently used the device’s camera to capture handwritten physical paper notes into the iPad — the interface for doing this, despite Evernote supporting multiple image files being store as one record, makes it difficult for users to capture multiple related images into one document. P19 again provides an exemplary comment: “If I’ve scribbled all over my meeting notes I want to be able to take snaps of each page in say, a 10 page document, and store them all together. Evernote doesn’t really make it easy to do that but I do it anyway and then store them in a folder or tag them so that I know they’re related. I can’t link them together without additional effort, so I just don’t bother.” This avenue may prove fruitful to future investigation. This theme links with the organisation and maintenance stages of the PIM framework — linking records together appears to be a significant shortfall of the current range of available software for iPad-style tablets, including Evernote, for the majority of our participants.

The final of the three themes that surfaced repeatedly during this initial analysis was that of tagging fatigue. Participants reported being stressed over the implication, as presented in the interface of the PIM software, that they needed to provide tags (keywords) for each record, feeling as though the software would make it difficult for them to locate the records again if they didn’t tag. Many past studies, such as Civan, Jones, Klasnja, and Bruce (2008), have discussed the experiences of users in tagging. P12 provided a comments that was typical of participants’ perspective on tagging in Evernote: “I hate it. The UI makes me feel like I have to tag everything, and that I won’t be able to find it again if I don’t, but really search just works fine without tags applied and the continual appearance of tags, or the lack thereof, in the interface just makes me feel like I have more work to do to keep things organised.” P14 also offers: “Tagging is so much extra effort, I have to think about my note to much more, what I want to use it for in the future, things like that, that I often just don’t want to tag it. I really don’t care, or know, if it [tagging] will even help me find it faster in the future.” We find this an interesting contrast from current scholarship in relation to tagging; recent work suggests that tagging might be the optimal solution for organising and refining information for certain groups of users (Voit, Andrews, & Slany, 2012). It is often suggested that tagging, as an act of categorisation, causes an individual to put more thought into the purpose and contents of the item being tagged (Craik & Lockhart, 1972; Jones & Teevan, 2003; Jones, Anderson, & Whittaker, 2012). Here, our findings suggest that individuals are highly aware of this additional mental overhead required from tagging and, correctly or otherwise, assume that the overhead is not worth the effort for potential benefits in future refining. This theme relates to all the stages of the PIM framework and suggests that a reappraisal of the role of tagging in, at the very least, iPad PIM software, may be worthwhile.

Discussion

While PIM is rapidly becoming a highly-studied field, the use of iPad-style tablet computers — a recent emergence in the consumer world — for PIM is under-studied and represents a significant area of interest to information science, human-computer interaction and information retrieval. The findings here underscore the conception that techniques for managing PIM collections differ across different technologies (such as laptop computers, desktops and iPad-style tablet computers), and the strategies that individuals use are affected not only by the choice of software, but by the hardware features and form factor of the device the software is used on. Prior investigations of how people utilise PIM software tools have largely been focused on traditional desktop operating systems.
The three themes identified above demonstrate that extant PIM software for iPad-style tablet computers is frequently failing to address, at the very least, the PIM needs of information workers. Data input, information linking and tagging are all core to the PIM experience, and while we chose to present the research through the lens of (the relevant components of) Barreau (1995) framework in this paper, other frameworks, such as Lin, Lutters, and Kim (2004), also identify these themes as equally important to PIM.

**Implications for Future Work**

From these findings, we can offer several suggestions for improved design of PIM software on iPad-style tablet computers, including the development of more nuanced options for image capture using the cameras, improved linking capability between existing and new notes, and improved, or removed, tagging interfaces to allow users flexibility in their choice to tag. Existing applications fail to meet many needs of the users discussed here, resulting in a variety of hacked improvisations.

In future publications, the component of the research presented here will also be situated in the results of the broader PIM study that the data collection was conducted as a component of.

**References**


Learning with Social Media: Measurement Tools for Understanding Information Behavior in Technology Pervasive Environments of the 21st Century

Leila A. Mills
University of North Texas
Leilamills01@gmail.com

Gerald A. Knezek
University of North Texas
gknezek@gmail.com

Jenny S. Wakefield
University of North Texas
jenny@wakefield.net

Abstract

This research reports findings from a study on information behavior for technology pervasive information environments in the 21st century. Social media users \((n=147)\) completed an online Learning Preference survey battery that included the Social Media Learning (SML) scale, the Technology Affinity Survey (TAS), the Computer Attitude Questionnaire (CAQ), and the Information and Communications Technology Learning (ICTL) survey. Findings revealed that 23\% of the variance in information seeking behavior, as measured by the Information and Communications Technology Learning survey, can be explained by a linear regression model including the SML scale, creativity and school attitude scales (CAQ), and TAS. Participants with higher ICTL scores for Information Seeking had greater preference for learning with social media, more positive attitudes toward school, higher self-reported creative tendencies, and lower preferences for immersive/always-on attachments to, or affinity for, modern information and communication technologies. Implications of these findings and future research directions are discussed.

Keywords: social media tools, guided inquiry, information seeking, information sharing, technology affinity

Introduction

Information and communications technology (ICT) tools play a central, commonplace role in formal and informal inquiry for information seeking, content sharing, and self-expression in technology pervasive information environments of the 21st century. While Web 2.0 technologies provide a wide array of ICT tools (including social media) that may be applied to information seeking for knowledge construction, educators and learning technologists are aware that simply having access to an array of content and contacts within a few clicks and keystrokes will not necessarily result in the use of ICT and social media tools for educationally relevant activities. The popularity of social media tools and the great interest in their potential for use in education prompted the research team to examine higher education students’ dispositions toward learning with social media tools, perceptions of creative tendency, attitudes toward school, and self-reported daily technology use in order to better understand information behavior. Survey data were gathered from an online learning preference survey battery that included four instruments: the Social Media Learning (SML) scale, the Technology Affinity Survey (TAS), the Computer Attitude Questionnaire (CAQ), and the Information and Communications Technology Learning (ICTL) survey, in order to determine the extent to which we can understand how ICT/social media tools may be used by individual students for information seeking. Such activities are considered educationally relevant and supportive of knowledge construction.

doi:10.9776/13286
Copyright is held by the authors.
Conceptual Rationale

Cognition

Cognition, social development, and technology. Vygotsky’s theory of Social Development recognized social interaction as a precursor to development, consciousness, and cognition in a progressive cognitive growth model where each function appeared in two forms: initially on a social level, and subsequently on an individual level (Vygotsky, 1962, 1978). Tharp and Gallimore (1988) described related foundation concepts from the work of social scientists in the neo-Vygotskian contextualist and interactionist school of thought. Scholars aligned with this school of thought believe that teaching, learning, and schooling can best be understood in a social context (Tharp & Gallimore, 1988), and that higher order cognitive functions develop from social interactions (Bruner, 1962). Bruner (1964) connected the development of cognition and the evolution of specialized human capabilities with technology tools. He theorized that, over time, humanity “has changed by linking himself with new, external implementation systems rather than by any conspicuous change in morphology” (Bruner, 1964, p.68). We see today many examples of ICT tools such as social media being used to augment information seeking and sharing capabilities.

Seeking Information

The information search process. Kuhlthau (1991, 2007) conducted research on student activities for affective, cognitive, and physical dimensions of information behavior in the traditional library setting, and also in technology rich online environments. The Information Search Process (ISP) Model, devised by Kuhlthau, depicts six stages of student information activity: initiation, selection, exploration, formulation, collection, and presentation. Together these six stages can assist to explain student information behavior and also allow educators to guide students in search activities (Kuhlthau, 1991). Kuhlthau’s (2007) research, an updated review of the literature and extensive inquiry project among n=574 students, indicated that the ISP model for dimensions of information behavior continues to be useful in explaining the search process in the digital, technology pervasive information environment of the Web 2.0 world (Kuhlthau, 2007; Kuhlthau, Heinström, & Todd, 2008).

Guided inquiry. Kuhlthau’s (1991) ISP model supports guided inquiry (Kuhlthau, Maniotes, & Caspari, 2007). Viewed as an instructional model, guided inquiry can direct students’ information behavior to educationally meaningful activities that will support knowledge construction. One approach to guided inquiry focuses on establishing a connection to each individual student’s learning environment. Maniotes (2005) conceptualized three spaces that are important for teaching and learning. The first space, she saw as the students’ current experience and knowledge, the second space as the curriculum, and the third space as the unique learning environment of a student. Maniotes’ model is based on the premise that educational social interaction (intellectual discourse) in the students’ third space can interconnect a student’s experience and knowledge (first space) with curriculum (second space) (Kuhlthau, Maniotes, & Caspari, 2007). Beyond supporting information behavior in a digital information setting, social media tools provide unique communication features that can support social dimensions of teaching and learning such as third space intellectual discourse.

METHODOLOGY

Research Questions

1. How do preferences for use of ICT social media tools and other technologies relate to student information behavior?
2. To what extent can we explain a tendency towards educationally relevant use of ICT from ICT social media tool and technology preference data?

Data Collection

Subjects were volunteer, higher-education, social media users, at least 18 years of age, who responded to email and Facebook invitations to complete the survey. The online, Learning Preference Survey Battery included four instruments: the Social Media Learning (SML) scale, the Technology Affinity Survey (TAS), the Computer Attitude Questionnaire (CAQ), and the Information and Communications
Technology Learning (ICTL) survey. This survey battery was selected and posted by graduate students enrolled in a North Texas (USA) learning technologies program to measure participants’ dispositions toward learning with ICT social media tools, perceptions of Creative Tendency, Attitudes Toward School, and self-reported daily technology use. One hundred forty-seven \( (n=147) \) subjects completed the survey during the spring semester of 2012. Subjects were 76% women \( (n=112) \) and 24% men \( (n=35) \) (Table 1), spanning 18 to 60+ years of age (Table 2).

### Table 1
**Descriptive statistics: study subjects by gender.**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>35</td>
<td>23.8</td>
</tr>
<tr>
<td>Female</td>
<td>112</td>
<td>76.2</td>
</tr>
</tbody>
</table>

### Table 2
**Descriptive statistics: study subjects by age.**

<table>
<thead>
<tr>
<th>Age in years</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>18-20</td>
<td>37</td>
<td>25.2</td>
</tr>
<tr>
<td>21-30</td>
<td>38</td>
<td>25.9</td>
</tr>
<tr>
<td>31-40</td>
<td>16</td>
<td>10.9</td>
</tr>
<tr>
<td>41-50</td>
<td>17</td>
<td>11.6</td>
</tr>
<tr>
<td>51-60</td>
<td>28</td>
<td>19.0</td>
</tr>
<tr>
<td>61+</td>
<td>11</td>
<td>7.5</td>
</tr>
<tr>
<td>Total</td>
<td>147</td>
<td>100.0</td>
</tr>
</tbody>
</table>

### Instruments

The Learning Preference survey battery measurement scales from CAQ, ICTL, SML, and TAS are comprised of questions with Likert-type response choices varying from 1 = *strongly disagree* to 5 = *strongly agree*. Cronbach’s alpha internal consistency reliabilities for all measurement scales were analyzed and interpreted by DeVellis’ (1991) guidelines (Table 3).

### Table 3
**Cronbach’s Alpha internal consistency reliabilities for instrument scales.**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Number of items</th>
<th>Item numbers</th>
<th>Alpha</th>
<th>Rating by DeVellis</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAQ: Creative Tendency</td>
<td>13</td>
<td>1-13</td>
<td>.78</td>
<td>Respectable</td>
</tr>
<tr>
<td>CAQ: Attitude Toward School</td>
<td>7</td>
<td>1-7 (2r,5r)</td>
<td>.81</td>
<td>Very Good</td>
</tr>
<tr>
<td>ICTL Total Scale</td>
<td>15</td>
<td>1-15</td>
<td>.77</td>
<td>Respectable</td>
</tr>
<tr>
<td>ICTL: Info Seeking</td>
<td>7</td>
<td>1,4,7,8,10,13,14</td>
<td>.71</td>
<td>Respectable</td>
</tr>
<tr>
<td>ICTL: Info Sharing</td>
<td>8</td>
<td>2,3,5,6,9,11,12,15</td>
<td>.83</td>
<td>Very Good</td>
</tr>
<tr>
<td>SML</td>
<td>7</td>
<td>1-7</td>
<td>.74</td>
<td>Respectable</td>
</tr>
<tr>
<td>TAS Total Scale</td>
<td>22</td>
<td>1-22</td>
<td>.74</td>
<td>Respectable</td>
</tr>
<tr>
<td>TAS: Immersed</td>
<td>13</td>
<td>7,10,9,8,13,12,3,21,22,11,4,16,19,5</td>
<td>.78</td>
<td>Respectable</td>
</tr>
<tr>
<td>TAS: AlwaysOn</td>
<td>4</td>
<td>18r,5r,2r,1r,6</td>
<td>.55</td>
<td>Unacceptable</td>
</tr>
</tbody>
</table>

**Computer Attitude Questionnaire (CAQ).** The CAQ was developed to measure attitudes toward learning and computers. The CAQ has foundations in earlier work funded by the Fulbright Foundation of Washington, D.C., the Japan Society for the Promotion of Science, and the Texas Center
for Educational Technology. The CAQ was formalized as a validated measurement tool in 1995 and has been extensively utilized in research studies (Knezek & Christensen, 1995, 2000) before being released for public use in 2000. The instrument was revalidated in 2011 (Mills, Wakefield, Najmi, Surface, Christensen, & Knezek, 2011). Internal consistency reliability for the CAQ Creativity scale ($\alpha = .78$) and the Student Attitudes toward School scale ($\alpha = .81$) for the 147 subjects in the current study were found to be respectable and very good.

**Information and Communications Technology Learning survey (ICTL).** The ICTL was designed and refined in a graduate course on psychometric measurement, and was expanded and validated in a 2011 study of student technology tool use (Mills & Knezek, 2012) (Figure 1). ICTL was developed to allow research on how students choose to interact with ICT tools in relation to educational information seeking and sharing. The refinement process for this instrument revealed two reliable measurement scales resulting from higher-order factor analysis (Dunn-Rankin, Knezek, Wallace, & Zhang, 2004): Information Seeking ($\alpha = .71$) and Online Information Sharing ($\alpha = .83$) with respectable and very good measurement properties, respectively.

**Information and Communications Technology Learning (ICTL)**

1. I would like to be a participating member of an online community.
2. I use Internet technology to explore topics of interest.
3. I like to share interests and reflections online.
4. I like to enroll in classes to continue my education.
5. I use Internet communications and other technology tools for self-expression.
6. I learn many things by interacting with other Internet users.
7. I like to take classes from good professors.
8. I use Internet communications technology tools when I want to learn about something new.
9. I learn best in a traditional classroom setting. (R)
10. Internet technology helps me be successful in my college classes.
11. More classroom learning should include interactive communication technology experiences.
12. The things I need to know are taught by instructors in the classroom.
13. I learn more when I regulate my own learning experience and seek information on things that I want to learn about.
14. I use Internet communications technology to keep current on topics related to my field of expertise.
15. I post information that might be of interest to other people.

*Figure 1.* The Information and Communications Technology (TAS) survey items.


**Social Media Learning scale (SML).** The SML was originally developed to measure student perceptions of Twitter for student reflections and community building in university courses featuring Global Policy and Digital Textuality. Subsequently the SML instrument was analyzed (Knezek, Mills, Wakefield, 2012) and refined by college faculty and learning technologies graduate students as the Social Media Learning (SML) scale (Figure 2). The refinement process revealed that this unidimensional (single factor) instrument has respectable internal consistency reliability ($\alpha = .74$).

**Social Media Learning (SML)**

When using social media.....

1. I feel a sense of community learning becomes interactive
2. Posting questions to my peers helps me understand my readings better
3. I am able to get faster feedback from my peers
4. I am able to get faster feedback from my instructor
5. I am able to communicate effectively
6. I am able to connect with peers more easily than face-to-face
7. I increase my participation in classes when I am allowed to contribute through social media

*Figure 2.* Social Media Learning (SML) scale items.

Technology Affinity Scale (TAS). TAS development was led by the first author during a doctoral level psychometrics class completed during summer 2011 (Figure 3). This instrument was inspired by the need for a reliable instrument to measure Internet related digital technology use — affinity for technology and immersive technology use – especially focusing on mobile technology tools. TAS produced high internal consistency reliability values as a total scale score (22 items) for the n=147 subjects in this study, and is currently being further refined to determine eventual retention of factors. Two measurement scales with established high construct (factor analytic) validity that are of interest in the current study include: preference for ubiquitous (always-on) communications (tas_AlwaysOn, \( \alpha = .55 \)), and tendency to be preoccupied with or immersed in technology-based interaction (tas_Immersed, \( \alpha = .77 \)) were found to be unacceptable and respectable.

<table>
<thead>
<tr>
<th>Technology Affinity Scale (TAS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. It is impolite to work on a computer in the audience during a presentation.</td>
</tr>
<tr>
<td>2. There are certain events during which ALL electronic devices should put away</td>
</tr>
<tr>
<td>3. My attention is often distracted by email or text messages when I am talking to someone.</td>
</tr>
<tr>
<td>4. I communicate with my friends mostly by text message</td>
</tr>
<tr>
<td>5. Some people are too absorbed in electronic communications to really listen face to face.</td>
</tr>
<tr>
<td>6. It’s okay to send text messages while carrying on a face to face conversation.</td>
</tr>
<tr>
<td>7. I often type text messages while walking down the street.</td>
</tr>
<tr>
<td>8. I sometimes check text messages while driving.</td>
</tr>
<tr>
<td>9. I sometimes check email messages during meetings.</td>
</tr>
<tr>
<td>10. I feel agitated when I am away from the internet for more than a day</td>
</tr>
<tr>
<td>11. I feel disturbed if I go out and forget my cell phone.</td>
</tr>
<tr>
<td>12. I prefer to socialize on social media rather than face to face.</td>
</tr>
<tr>
<td>13. Many relationships are easier to maintain on facebook-type social media.</td>
</tr>
<tr>
<td>14. I would use an online dating service.</td>
</tr>
<tr>
<td>15. I would not use the internet to find a babysitter.</td>
</tr>
<tr>
<td>16. My computer is just as important to me as my wallet or purse.</td>
</tr>
<tr>
<td>17. For me, a computer is a better companion than a pet.</td>
</tr>
<tr>
<td>18. Many people are too attached to their smart phones.</td>
</tr>
<tr>
<td>19. Many people have good friends they met via social networks.</td>
</tr>
<tr>
<td>20. Getting married via computer connection is taking the internet a bit too far.</td>
</tr>
<tr>
<td>21. Sometimes I feel more available to my electronic devices than to my family</td>
</tr>
<tr>
<td>22. I sometimes feel I am a slave to the technologies that surround me.</td>
</tr>
</tbody>
</table>

Figure 3. Technology Affinity Scale (TAS) items.
Note: TAS V2.0 by G. Knezek and L. Mills 12/2011

Data Analysis and Findings

Correlation Analysis. Identifying and understanding how information seeking behavior correlates with technology preferences was of particular interest in this research. A review of the ICTL Information Seeking scale items revealed content of possible usefulness for the measurement of constructivist information acquisition behavior that would be indicative of knowledge construction for learning. Pearson product moment correlation analysis was conducted in order to examine possible relationships between ICTL Information Seeking values and values for social media, technology, and CAQ learner dispositions. Information seeking was found to be significantly (p< .05), positively correlated with learning through social media, SML (r= .34, p< .0005); CAQ Creative Tendencies (r= .25, p< .002), and CAQ Attitudes Toward School (r= .20, p< .017).

Analysis of variance for high versus low dispositions. An analysis of variance was conducted for low vs. high affinity for learning with ICT (partitioned based on ICTL Total Scale Score), in order to determine if the two groups would have significant mean differences for SML, TAS, and scales of the CAQ (Creative Tendencies, Attitudes Toward School). A low/high ICTL designation was assigned based on the ICTL Total median score, with the median score for n=147 found to be 3.69 of possible 1-5. Respondents with mean ratings below the median were designated as low, while those with mean ratings above the median were designated as high. Significant differences (p<.05) for low vs. high ICTL values were found for SML, TAS immersion, and CAQ Creative Tendencies. To summarize (Table 4), students in
the group found to have high ICTL information seeking reported a more positive disposition for learning with social media, tended to be technology immersed, and felt creative as learners. Effect sizes for these significant discriminations were in the large (ES = .8) to medium (ES = .5) range according to guidelines by Cohen (1988).

Table 4
Analysis of Variance for high versus low means for ICTL Information Sharing groups.

<table>
<thead>
<tr>
<th>Scale</th>
<th>N=</th>
<th>Mean</th>
<th>N=</th>
<th>Mean</th>
<th>Sig.</th>
<th>Cohen’s D Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low Group 1</td>
<td>High Group 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SML</td>
<td>75</td>
<td>3.10</td>
<td>72</td>
<td>3.61</td>
<td>.000</td>
<td>.86</td>
</tr>
<tr>
<td>TAS: Immersed</td>
<td>75</td>
<td>2.87</td>
<td>72</td>
<td>3.21</td>
<td>.001</td>
<td>.56</td>
</tr>
<tr>
<td>CAQ: Creative Tendencies</td>
<td>75</td>
<td>3.77</td>
<td>72</td>
<td>4.00</td>
<td>.002</td>
<td>.51</td>
</tr>
</tbody>
</table>

Low vs. high group analysis of variance was also performed for the ICTL Information Seeking scale. The median value for this scale for n=147 subjects, which was used to separate the low group versus the high group, was 4.2 of a possible 5.0. Group mean differences with educationally meaningful effect sizes (Cohen, 1988; Sivin-Kachala, Bialo & Langford, 1997) were found for Social Media Learning (SML), Technology Affinity Immersion (always on) connected communications (TAS_alwaysOn), CAQ Attitudes Toward School, and Creative Tendencies. Participants with higher ICTL scores for Information Seeking had greater preference for learning with social media, more positive attitudes toward school, higher self-reported creative tendencies, and lower preferences for immersive/always-on attachments to or affinity for modern information and communication technologies. Effect sizes for group mean differences were in the small to moderate range according to guidelines by Cohen (1988) (Table 5).

Table 5
Analysis of Variance for high versus low means for ICTL Information Seeking groups.

<table>
<thead>
<tr>
<th>Scale</th>
<th>N=</th>
<th>Mean</th>
<th>N=</th>
<th>Mean</th>
<th>Sig.</th>
<th>Cohen’s D Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low Group 1</td>
<td>High Group 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SML</td>
<td>73</td>
<td>3.24</td>
<td>74</td>
<td>3.46</td>
<td>.022</td>
<td>Toward.37</td>
</tr>
<tr>
<td>TAS: Always-On</td>
<td>73</td>
<td>1.93</td>
<td>74</td>
<td>1.75</td>
<td>.052</td>
<td>-.33</td>
</tr>
<tr>
<td>CAQ: Attitudes Toward School</td>
<td>73</td>
<td>3.46</td>
<td>74</td>
<td>3.70</td>
<td>.046</td>
<td>.32</td>
</tr>
<tr>
<td>CAQ: Creative Tendencies</td>
<td>73</td>
<td>3.82</td>
<td>74</td>
<td>3.95</td>
<td>.066</td>
<td>.29</td>
</tr>
</tbody>
</table>

Regression Analysis. Regression analysis was used to determine the extent to which information seeking behavior could be explained by other perceptions measured in the study. A linear combination of learning with social media (SML), TAS for always-on connectedness, CAQ Attitudes Toward School and Creative Tendencies accounted for 23 % (RSQ=.23) of the variance in ICTL Information Seeking behavior (p<.0005). The standardized regression coefficients (Beta weights) indicate that preference for SML (Beta = .364, p< .0005), and CAQ Attitudes Towards School (Beta = .269, p =
.003) best account for the variance in Information Seeking in this multiple linear regression model (Table 6).

Table 6
Linear regression standardized regression coefficients and model summary.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>1.1914</td>
<td>.451</td>
</tr>
<tr>
<td>TAS: Always-On</td>
<td>-.150</td>
<td>-.152</td>
</tr>
<tr>
<td>CAQ: Attitudes Toward School</td>
<td>.170</td>
<td>.229</td>
</tr>
<tr>
<td>CAQ Creative Tendencies</td>
<td>.190</td>
<td>.156</td>
</tr>
<tr>
<td>SML</td>
<td>.335</td>
<td>.364</td>
</tr>
</tbody>
</table>

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>RSQ</th>
<th>Adjusted RSQ</th>
<th>Std. Error of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.476*</td>
<td>.226</td>
<td>.205</td>
<td>.4884</td>
</tr>
</tbody>
</table>

Note: a. Dependent Variable: ICTL_Seeking
Note: a. Predictors: (Constant), SML, TAS_AlwaysOn, Creative Tendencies, Attitudes Toward School

Discussion

Technology pervasive information environments offer unique affordances for educational discourse and information seeking behavior that is associated with social and cognitive development for inquiry and knowledge acquisition. The interactive Web 2.0 environment provides new communication tools that are easily mastered and maneuvered by most students. Yet as an increasing number of daily activities and educational offerings are transferred to the online realm, educators must be able to guide inquiries and provide interventions to support learning. Reliable measurement tools can assist in gauging the quality and quantity of ICT tool information behaviors and provide insight for directing information behavior in the direction of creative inquiry and discourse that will support construction of knowledge. Further research is needed to firmly establish the reliability and validity of these and additional instruments that can support efforts to create learning opportunities that maximize the affordances of ICT and social media tools for the construction of social-mediated knowledge. This approach is consistent with the theoretical conceptions of Vygotsky (1978) and Bruner (1969) when social media is used in a manner that creates opportunities for cognitive functions to appear on a social level (Vygotsky, 1962, 1978), for teaching and learning in a social context (Tharp & Gallimore, 1988). Social media tools may augment individual cognitive functions (Bruner, 1962), and provide communication options that can connect curriculum and teacher direction to the realm of student experience and knowledge (Kuhlthau, Maniotes, & Caspari, 2007). Additional research is planned to verify the direct application of these theoretical constructs to social media-assisted learning in a social context.

Conclusion

This research reported findings from a study of information behavior in technology pervasive educational environments of the 21st century. Social media users (n=147) completed an online learning preference survey battery that included instruments designed to understand student information behavior for seeking and sharing information in technology rich educational environments. Preference for learning with information and communications technology was assessed as a function of the other measures.
Major findings were that students with total ICTL scores for the subscale Information Seeking had greater preference for learning with social media, more positive attitudes toward school, higher self-reported creative tendencies, and lower preferences for immersive/always-on attachments to or affinity for modern information and communication technologies. Additional findings revealed that 23% of the variance in information seeking behavior, as measured by the Information and Communications Technology Learning survey, can be explained by a linear regression model based on instrument scales from SML, TAS, and CAQ. Further research is required to refine knowledge of specific behaviors associated with information seeking preferences in a technology-rich educational environment. Additional studies of information seeking predictive models are planned for the future.

References

Visualizing Overlapping Latent Communities Using POI-Based Visualizations

Patrick M. Dudas  
University of Pittsburgh  
pmd18@pitt.edu

Jae-wook Ahn  
Drexel University  
ja626@drexel.edu

Martijn de Jongh  
University of Pittsburgh  
m.a.dejongh@gmail.com

Peter Brusilovsky  
University of Pittsburgh  
peterb@pitt.edu

Abstract

Social network analysis and social network visualizations can provide a meaningful statistical and topological understanding of latent communities. However, the majority of current visualization approaches just represent sub-communities as clusters of closely related nodes in a node-link diagram and embed limitations to represent overlapping communities and multi-layer community structure frequently found from modern complex networks. We argue that visualizations based on points of interest can provide a better solution to represent overlapping latent sub-communities. We present two visualization systems, SuperVIBE and ContextForces, which implement this approach. These systems operate by creating two-dimensional latent spaces by means of grouping nodes using external variables not presented in the graph and by offering an interactive visualization to filter and map in these latent spaces. Understanding which latent groups are most central to a variety of topics and providing visual clues to the individuals critical to those groups provides a mechanism to explore and discover overlapping latent communities.

Keywords: social network analysis, latent community, visualization, SuperVIBE, ContextForces

Introduction

In recent years, social network analysis and visualization have emerged into a very popular topic (Fortunato, 2010). A reasonable fraction of research on this topic focuses on visualizing a set of latent (sub)communities within the rich structure formed by social connections among people. It is interesting, however, that the last 10 years brought almost no changes to the dominant approach of presenting the structure of discovered communities: the majority of modern work still present sub-communities as clusters of closely located nodes in a traditional node-link diagram (Newman & Girvan, 2004; Wakita & Tsurumi, 2007). The only major change – a gradual move from static to interactive (exploratory) social network analysis (SNA) that offered users an ability to interact and manipulate the graph or statistical variables to highlight key nodes or leaders (Perer & Shneiderman, 2008) – has not augmented or improved the visualization approach itself.

Social networks explored by modern researchers are typically much more complex than simple node-link diagrams. These “modern” social networks are based on larger and more heterogeneous data (frequently extracted from online systems) and frequently represent several different dimensions of similarity among people. To deal with this increased complexity and heterogeneity, data mining researchers introduced more and more sophisticated community-mining approaches. A number of these approaches were specifically created to discover overlapping communities and multi-layer community structure. In this context, the dominated node-link visualization approach created for presenting simpler non-overlapping communities emerged as a bottleneck, restricting our ability to analyze and comprehend modern complex community structures.


Copyright is held by the authors.
In this paper we suggest two alternative approaches to visualizing complex, overlapping or multi-layered community structures based on interactive exploration with Points of Interest (POI). The POI-based visualization was originally created to visualize search results; however, we believe that with some moderate modification, this technology can provide a great approach for exploring (overlapping) latent communities. The paper presents two POI-based approaches for exploring latent communities: (1) SuperVIBE introduces a few changes to the original POI-based visualization known as VIBE; and (2) ContextForces attempts to bridge the gap between traditional node-link diagrams and a POI based visualization. The following section describes, in detail, the two visualizations, and concluding section looks at the future work and development of both of these projects.

Dataset

Network

To explore multi-layer social network visualizations, we used a dataset containing several different types of connections among a group of authors that have published in the UMAP (User Modeling, Adaptation and Personalization) conference series. The UMAP is the main conference series on adaptive systems. To create this dataset, we have extracted connection data from the DBLP database (Ley, 2002) and the Conference Navigator 3 (CN3) system. CN3 was developed at the University of Pittsburgh’s iSchool (Parra, Jeng, Brusilovsky, López, & Sahebi, 2011) and it supports conference organizers and delegates with web-based informational, scheduling, and social features. Social features let the users see each other’s schedules and connect using CN3 networking options (make a bidirectional connection or follow somebody). It also provides third party social network information, including Facebook, LinkedIn, CiteULike, BibSonomy, and Mendeley. CN3 has been used at several UMAP conferences and we extracted co-bookmarking data to socially connect authors, if they bookmarked similar sets of papers within the CN3 system. We also used the author’s publication data in order to connect them if they published articles with similar topics. The social networks were created from the data by aggregating co-authorship information and similarity measures into links, and storing authors as nodes.

Latent Communities as POIs

The main challenge of any POI-based approach is the selection of meaningful POIs. While there are many possible options for POI selection in the context of multi-layer communities, we started our exploration of multi-layer community visualization with a specific POI selection approach that we consider as most promising and universal. This approach is based on two principal ideas: (1) the POIs used for visualization correspond to the top 10 or 20 latent communities extracted from a multi-layer social network; (2) the latent communities (i.e., POIs) are discovered using Latent Dirichlet Allocation (LDA) algorithm (Blei & Lafferty, 2007). LDA is an algorithm widely used for probabilistically discovering topics from a set of documents based on keyword frequencies.

Following this approach, we applied LDA to the multi-layer data formed by UMAP authors’ publication or bookmarking activities in order to elicit latent communities represented as research topics. The process is as follows:

(1) Extract papers the authors bookmarked or published;
(2) Convert the papers to a bag-of-word representation (author-keyword matrix);
(3) Feed the matrix to LDA;
(4) Two datasets (bookmarking and publication) are created using LDA. Each includes two outputs:
   (a) topic-probability pairs per each author; (b) list of keywords with higher probabilities per each topic.

Output (a) maps the authors and the topics (latent communities represented as POIs) and (b) provides the keyword of each community (e.g. “social, search, web” in Figure 1).

602
POI-Based Overlapping Latent Community Visualizations

The social network community discovery methods mostly rely on graph structures by dividing network nodes into densely connected subgroups or clusters (Newman & Girvan, 2004; Wakita & Tsurumi, 2007). They are more appropriate for representing one community for each person in the network and cannot support overlapping latent communities, where community membership is not decided by explicit conditions (latent) and a single person can participate in multiple communities (overlapping membership). An example of these types of communities is an academic conference, where people gather and form latent communities following their research interests but these communities may not be consistent with their physical affiliations. Rather, their activities such as co-bookmarking behaviors within a conference management system can better aid such community discovery task. The membership derived from these activities is not explicitly defined and one person can be a member of multiple communities. Our SuperVIBE and ContextForces systems are devised to best represent this type of community. They are expected to meet the following requirements:

1. Visualize multiple latent community membership of a single person.
2. Visualize the different degree of membership. One person can be more strongly related to one community than the others.

Latent communities represented as POIs SuperVIBE and ContextForces can fulfill these requirements. People are placed closer or further from the multiple POIs (Requirement 1), according to their degree of membership (Requirement 2). Additionally, ContextForces keeps intact the node-link paradigm to add an additional layer of network connectivity. The details are introduced below.

SuperVIBE

VIBE (Visual Information Browsing Environment) is a POI-based visualization developed by (Olsen, Korfhage, Sochats, Spring, & Williams, 1993). It displays the POIs and related objects by locating the objects according to the similarity ratio to the POIs. An object that is more similar to a POI is placed closer to the POI and the distances to POIs are always consistent with the similarity ratios. VIBE can visualize N-to-N relationships (objects-to-POIs) and it makes VIBE suitable for visualizing overlapping people-to-latent communities relationships. We expect VIBE can benefit the following tasks:

1. Discover and rank the top N members for overlapping latent communities.
2. Discover which communities are more relevant for a member.

Figure 1 shows the relationships between the latent communities (POIs in yellow circles) and the authors (squares) in VIBE. The POIs are labeled with latent community concepts discovered by LDA using the authors' co-bookmarking actions. Despite the straightforwardness of the VIBE algorithm, its one shortcoming is that it is based on similarity ratios and easily creates clutter when the differences between the similarity value ratios are smaller (Figure 1). Therefore, we extended the traditional VIBE as SuperVIBE by modifying the visualization algorithm as follows:

1. Consider only the POIs that have the similarity values greater than a specified threshold.
2. Consider only the top N POIs for a user.
Figure 1. VIBE visualization of co-bookmarking communities. POIs (circles) are communities labeled with key-concepts and squares are authors. Most are cluttered in the center.

These features reduce the influences of the POIs with lower membership similarity and make the users move towards the ones with higher similarities. We allow SuperVIBE users to switch between the two features, in order to discover the best visualization ("Controls" in Figure 2 & 3). Figure 2 shows that it can reduce the clutter in the center by considering the POIs that have similarities greater than a threshold (0.05).

Figure 2. SuperVIBE using the "similarity threshold" method

Figure 3 shows a more revealing visualization. This time, it considers the top N POIs from each author when calculating their positions. It clearly shows several clusters gathered around the POIs. One big cluster is found just below the "SOCIAL SEARCH WEB" and "USER ANALYZE FEEDBACK" latent community POIs, which suggests that the authors with the interests in those topics belong to the overlapping latent communities. Another example is a cluster of five people that are gathered around
“MUSIC RECOMMEND TAG” and “DOMAIN COLLABOR FILTER.” We can assume that these authors are involved in music recommendation and collaborative filtering communities.

Figure 3. SuperVIBE Using Top N POI Method

It also supports visually discovering community members from POIs. Figure 4 shows an example of searching for people who might belong to the “MUSIC RECOMMEND TAG” community. When the mouse cursor is moved over a POI, the authors that are highly similar to the POI are highlighted. The size of the highlights (bigger, more similar) and the color (red, more similar) corresponds to the strength of a similarity. Therefore, we can see that the 4 people with big, red highlights are the most relevant members of the community. Even though the proximity in the visualization (POI to authors) can also support the task, this color-based method can help users complete the task more quickly.

Figure 4. Finding members from a community POI

ContextForces

ContextForces provides both the mechanism of POI-based visualization and traditional node-link diagram using force directed placement (FDP). It was designed using an open-source project called
D3.js (Bostock, Ogievetsky, & Heer, 2011). For examining the roles of actors in a social network, a plethora of the research has focused primarily on the development of groups, communities, cliques, and mutually-relevant, homogeneous clusters of individuals based on metadata about the actors (Ahn, Han, Kwak, Moon, & Jeong, 2007; Heer & Boyd, 2005; Matsuo et al., 2006). However, this is mostly tapered to statistical calculations or user-generated interpretation. ContextForces’ POI-based approach applies the properties of a Venn diagram (e.g. InfoCrystal (Spoerri, 1993)) to the baseline FDP. It relies on the FDP (Fruchterman & Reingold, 1991) framework to designate the placement of the nodes as the user interacts with the visualization. In a typical FDP environment, nodes are placed based on the edges and the weighting between each edge. For example, Figure 5 presents a synthesized FDP arrangement.

Figure 5. Synthesized FDP network

Figure 6 presents the same arrangement after POIs are applied and are locked to the apex of a circular shape, arranged in equal degrees from one another around the network. In this figure, nodes are augmented to reflect these variables and will render themselves to those new positions based on the weights placed by the POIs. The resulting position of the nodes is produced by taking into account both traditional FDP placement and POI attraction. To stress the impact of POIs, nodes can be viewed in a Venn-diagram like manner, highlighting the interaction of the external variables (Figure 7).

Figure 6. POIs pulling nodes from the FDP layout (based on weighted edges)
Researchers can apply any POIs to redefine the latent communities. ContextForces subdivides these very large networks (in the case of the UMAP dataset, 766 nodes and 19,000+ edges) into, at most, \((n^2 - n + 1)\) subdivisions.

The concept of ContextForces is to add POIs and allow the visualization to be altered interactively so researchers can see how the POIs and the network are unified. At the same time, users can switch between the POI-based visualization and FDP approach and see how the network evolves. In this case, Affinity Propagation clustering algorithm (Frey & Dueck, 2007) is applied to the network and the subsequent color represent the community (or cluster) of each node. We implemented functionality to either relax or exacerbate either of these two mechanisms, including the:

1) Ability to shrink or elongate the distance between the nodes within a cluster or nodes attached to POIs (Figure 8)
2) Ability to increase the font size based on linearly increasing the value or based on degree centrality of each node in the original topology (Figure 9)

Figure 8. Shrinking links between inner-cluster nodes and POIs and nodes
Lastly, we define edge weights for each author to each topic based on LDA values. We can then trim these edges to show better exemplar authors for each topic. Figure 10 shows the complete graph without trimming and Figure 11 shows the trimmed graph.

Figure 9. Degree centrality font

Figure 10. Complete graph and POIs
Discussed in the paper is a network based on the UMAP dataset and latent community structures determined by LDA. We propose two novel approaches to spatially visualizing the POIs representing these UMAP latent communities. SuperVIBE considers both the POIs that have the similarity values greater than a threshold and only the top N POIs for a user. ContextForces looks at these same two criteria, but also applies the node-link diagram using force directed placement. We then applied the UMAP dataset and the latent community structures determined by LDA to both visualizations to showcase a new paradigm in the visualization of social network analysis.

We expected that these two approaches could better define overlapping latent communities using spatial methods and promoting a new paradigm in mapping data with network visualization. We believe this mapped data is limitless and without bounds in terms of context in that any latent artifact can be mapped to each visualization and new insight can be found about network.

While we believe that the suggested POI-based approaches offers an important advantage to those exploring modern online communities, the true impact of this approach has to be determined in a user study. We will approach using comprehensive user-study to test whether our tools can achieve the expected goal and to identify which features allow for the identification of key individuals in a given network dataset in a timely fashion, among other network or overlapping latent community facets. We also plan to extract latent communities using additional social activities such as co-attending conferences. If this does warrant a valid hypothesis, we would then like to select other datasets that require identification of key players including: dark networks (i.e. suspected criminals or terrorists), medical data, conference authorship, and others. Utilizing these diverse datasets we will explore and develop novel visualization functions to further improve exploratory SNA.
References


The Atlas Project—Comparing Time and Place

Tim Nugent
University of Tasmania
tnugent@utas.edu.au

Christopher Lueg
University of Tasmania
Christopher.Lueg@utas.edu.au

Andrew Bennett
University of Tasmania
acb0@utas.edu.au

Abstract

The Atlas Project is an interdisciplinary project investigating spatial and historical exploration of geographical areas, with Hunters Wharf in Hobart, Australia being the example space used in this research. One of the goals of this research is to create a mobile application allowing historical information to be transferred out of the traditional static historical media and made more easily available for a variety of different tasks. Through a heavily iterative process, a prototype designed through continuous discussion amongst three different groups led to the research focussing on comparison of historical maps supported by additional spatiotemporal indexed features as the primary means of interacting with historical data. This paper outlines the iterative development of this prototype, presenting research challenges and solutions encountered during that process; it introduces a novel approach of presenting both spatially and temporally sensitive information that satisfies multiple workflows and multiple disciplines. Finally, the paper presents many interesting opportunities for future work and further analysis toward a generalised method of providing context sensitive spatiotemporal map overlays on mobile devices.

Keywords: spatiotemporal systems, historical maps, tourism, exploration, interdisciplinary, interaction, mobility

Introduction

The Atlas project is a multidisciplinary project between the University of Tasmania schools of History and Classics, Computing and Information Systems, Architecture and Engineering. Atlas is attempting to create a new means of exploring historical information, by combining historical data with up-to-date spatial maps and mobile interaction. The project aims to enable people from different domains and with different goals to gain a new insight into the historical data which has been traditionally locked up in static media such as paper maps and books. With so many varied disciplines working on a single project, each are likely going to have ideas for such a system that are incongruous with the other disciplines. One of the great challenges this project had to overcome was deciding which goals and features to include and which to modify as the research evolved over time. The primary objective of the project was to combine historical data, primarily in the form of high resolutions maps of different eras, with current spatial data. The primary means of gaining insight from this information was to be through comparison; mostly comparison between the historical maps and the current terrain. This led to the decision to create a mobile prototype system through rapid iteration to allow for the greatest flexibility amongst the differing goals of each school whilst still remaining in sync with the overall vision for the system. The idea of visualising historical events and data is certainly not unique to this research. (Friendly, 2008) document one of the earliest known visualisations of historical data, from the 18th century, their research depicts a gigantic plot of almost 3000 years of births and deaths of famous historical figures.

More recent technology has certainly aided the increase in historical visualisations, since the 90’s Geographic Information Systems (GIS) tools and techniques have been a popular choice for visualising historical data (Coppock & Rhind, 1991; Maguire, Goodchild, & Rhinds, 1991). Modern GIS tools such as Google Maps have allowed geographic historical visualisations to become more and more varied, such as the HistoryLines system by (Klaebe, Foth, Burgess, & Bilandzic, 2007) for the newly built Kelvin Grove Urban Village in Brisbane. HistoryLines allowed residents to easily see where other residents had lived in the past as a series of lines and points on a map. Historical visualisations have even started taking on

Copyright is held by the authors.
more artistic and social properties, such as the BBC Dimensions project, which allows for popular historical events, such as the path and distance the Mars rovers have taken, or the flooding in Pakistan, to be shown as an overlay anywhere on Earth (BBC, 2011). The historical art video ‘A History of the World in 100 seconds’ (Lloyd, 2011) was created by parsing all Wikipedia articles with location and date metadata. The video progresses through time from 387BC to the present day, flashing notable historical events at both the time and place they occurred in a world map. Whilst this is a very focused view of history, based solely on what Wikipedia had available, it is still a novel means of seeing a snapshot of the historical information available for general access. With a goal in mind for the system and what had been done before the next step was to choose a historically interesting area to use as a data source for the prototypes development.

**Hunters Wharf**

Hunters Wharf is an area of the Hobart city waterfront which has undergone a large amount of change in it’s relatively short lifespan of European settlement. Hunters Wharf was settled in 1804 by Europeans after failed settlement attempts in other areas of Tasmania. Originally an island the space between the island and rest of the city waterfront was eventually filled in. Nowadays there is no obvious signs of Hunters Wharf once being Hunter Island. Over time the development of Hobart city had the Hunters Wharf area goal of having a great deal of merchant and industrial facilities, including a gasworks, iceworks and a jam factory; many of which are now classified as historical buildings by the Hobart council. The Hunters Wharf area today is mostly a tourist site and component of the Hobart city waterfront area, as well as hosting the University of Tasmania School of Arts building in an old iceworks building. Despite it’s relatively young age, Hunters Wharf has had quite a busy life with a great deal of building and landscape development undertaken in this time. Due to this heavy activity Hunters Wharf was chosen to be the area that the Atlas project would initially focus upon. It was decided for this research to initially focus on a small, busy, area in greater detail rather than try to cover a larger area in less detail.

**Competing Goals**

As stated above, the Atlas project is an interdisciplinary research project between several University of Tasmania schools, each with their own views as to the best way for historical and spatial data to be combined and each with differing goals for the project to meet. The researchers from the Computing and Information Systems were focussed on the exploration aspects of the available data. The intention being that anyone would be able to use a mobile device and be able to easily see the changes that had been undertaken in the very area of the Hobart waterfront they were currently exploring. The researchers from History and Classics goal was to have a system allowing them to better visualise the historical data already available to them, in some ways an electronic and interactive encyclopaedia. They wanted a system where they could overlay numerous maps to illustrate changes, as well as support for spatiotemporally indexed points linked to supporting information. The School of Engineering wanted a system to that would assist with surveying and relating tasks when working near or on historical areas of Hunters Wharf, they wanted to be able to see at a glance if they were about to dig into historically significant positions, old piping, sewage, or foundations of systems that had been built up over the years. These are three very different goals yet they do have an important element in common; they all are based around the idea of seeing spatial changes over time. Combined with the majority of the historical data available in the form of maps, it became apparent that the most obvious method of continuing would be through maps overlaid over a current map of the Hunters Wharf area. This commonality amongst the dissonance led to an environment of constant discussion, compromise and feedback. Every two weeks each of the researchers would meet in person and discuss any changes to the project and the current iteration of the prototype. New features and changes to existing ones were debated and decided on during these meetings. After each meeting a new iteration of the prototype was developed and distributed to the members as rapidly as possible to allow them time to experiment with the prototype before the next meeting. The decision to develop for mobile platforms was made very early on with each party agreeing that any future system to come out of this research, regardless of the specific task it was being used for, it should be able to be used in the field to give context. As such any prototype needed to be mobile to facilitate this.

**Prototype evolution**

The iPad was chosen as the primary development platform as it possesses a large enough screen to better enable fine details to be seen in maps whilst still being easily portable. The iPad was also seen as a device that anyone interested in the Hunters Wharf area may be carrying regardless of their goals. Additionally, the iPad has a great deal of existing libraries and services for presenting and manipulating
geographic data and maps with ease. Finally the programmers for this research also had experience developing on the iPad platform. The advantage of development experience and the existence of iPad libraries was the main reason why the prototypes were developed as high-fidelity, (Rogers, Sharp, & Preece, 2003). It was not much more effort to create and experiment with a working proof of concept than it was to mock it up in any other form, allowing constant feedback with a working system at each stage of the project.

The rapid high-fidelity prototyping through pre-existing software is very similar in nature to web-mashups, allowing new ideas to come forth that are not immediately obvious without a working system to demonstrate them (Floyd, Jones, Rathi, & Twidale, 2007).

The prototype was initially just a single historical map overlaid onto a current map of the Hunters Wharf area, with a single slider at the bottom of the screen to adjust the opacity of different historical maps. Whilst at this stage the prototype was mostly a technical exercise, investigating how complex it would be to display and change the opacity of a map, it proved popular with the entire project team and became the base of the rest of the prototype evolutions.

The next significant evolution was a split view that allowed the user to select from one of several different historical maps, replacing the previous historical map being overlaid. This feature came from the desire to quickly contrast different time periods with how the region currently exists, as such there needed to be an easy way to switch between historical maps, similar to flipping between pages in a book to quickly compare information between two pages. The prototype was then extended so that two historical maps could be overlaid at once and both have their opacity changed independently of each other. This was a logical flow on from the ‘page-flipping’ idea of quickly changing maps, allowing for more advanced comparing and contrasting, such as having an newer map with the highest opacity possible to just see the modern layout of Hunters Wharf below an older map, with an even higher opacity, allowing for three different eras of the area to be seen at once; or for just comparing directly two historical maps without the need to be bothered with the modern view.

Figure 1. A historical map overlaid onto a current view of Hobart

An additional information overlay was then created to allow additional information to be displayed alongside the maps to provide extra information than what maps alone could provide. This overlay was hooked into the main map view as pins, geospatially placed at the location relevant to their information. The idea behind the additional information was that they would ultimately be tailored to the specific user,
such as interesting historical events for tourists or important historical construction details for builders.

Finally the prototype gained the ability to mark off and highlight specific regions on any of the maps, this was added in at the request of the engineering project members to make analysing an area easier, but it was also seen as a potentially useful feature for anyone to use.

The engineering members also requested a means to modify how the maps were added. Initially whenever a historical map was overlaid, it zoomed the view out to fit the entire map so that the user could see the area the historical map covered. When trying to compare very specific areas at higher zoom levels however, the zooming out behaviour forced the users to replace their view before they could continue comparing. As such this zooming feature was deactivated.

![Figure 2. Several highlighted regions of the map](image)

In the next iteration the prototype was ported to the iPhone to better extend the reach of the project. The iPhone version of Atlas is limited to viewing only one historical map at a time, as the two maps caused too much clutter on the smaller screen to be considered useful. The iPhone version also had a different, more paginated, style of navigating through the historical maps and information and did not have the ability to highlight areas. Again these changes were to limit the amount of clutter on the smaller screen and otherwise the iPhone and iPad versions were identical in features.

**Conclusions and the Future**

The Atlas project is an interdisciplinary project attempting to make historical information more accessible for a variety of different goals for exploration of the Hunters Wharf area of Hobart. As the project advanced it was found that comparison through overlaying historical maps allowed for easy comparison amongst the past and present and the past again. The next step for this work is to evaluate the system with users from the different disciplines. As with the preliminary prototype it is expected that maps overlaid with historic information will be well received by all interested parties. It is also expected to both improve productivity and offer opportunities for a unique and more widely accessible exploration of historic data sets within their modern context. This opens potential research avenues in History, GIS, data
visualisation and HCl, as well as opportunities in the Tourism sector. It is expected that future opportunities and challenges will present themselves as the prototype undergoes testing.

References


Regulating Anti-Social Behavior on the Internet: The Example of League of Legends

Yubo Kou
University of California, Irvine
yubok@uci.edu

Bonnie Nardi
University of California, Irvine
nardi@ics.uci.edu

Abstract

Anti-social behavior such as flaming is pervasive and problematic in many online venues. This behavior breaks established norms and unsettles the well-being and development of online communities. Therefore, regulating anti-social behavior becomes an important issue in online management. In this paper, we examine flaming in the context of an online game, League of Legends. We discuss an effort the game developer, Riot Games, presents to deal with anti-social behavior. The "Tribunal System" empowers players to judge misbehavior. We conducted an ethnographic study of the game to explore the reasons for anti-social behavior, and we analyzed the development of the Tribunal System.

Keywords: flaming, regulation, video game, League of Legends

Introduction

Anti-social behavior is a pervasive and real problem on the Internet (P. Davis, 2002). Researchers have suggested that the reduced social cues and social anonymity of computer-mediated communication result in anti-social behavior (Kiesler, Siegel, & McGuire, 1984). One type of online anti-social behavior is flaming which indicates aggressive, hostile, sometimes profanity-laced interaction. Flaming is widely found in email, public newsgroups, discussion boards, and online video games (O’Sullivan & Flanagin, 2003; Spears & Lea, 1992; Lea & Spears, 1991; Pizer, 2003; Postmes, Spears, & Lea, 2000; Lea, O’Shea, Fung, & Spears, 1992; Alonzo & Aiken, 2004; Thompisen, 1992).

To regulate people's online behavior, Lessig (2000) discussed four modalities: laws, norms, markets, and code. Laws regulate behavior by threatening a certain consequence if a law is broken. Norms regulate behavior by sanctions imposed by a community. Markets regulate behavior by pricing structures that constrain access. Code (i.e., software and hardware) regulates behavior by constituting a set of constraints on how people can behave. The constraints are experienced as conditions on people's access to cyberspace. For example, in some places people must enter a password to gain access. Lessig argued that laws are difficult to enforce on the Internet. So are markets due to the free nature of the Internet. The most common modalities people use are norms and code.

In this paper, we selected League of Legends to study online behavior because it is a widely played game and people must cooperate quickly with one another. A match is composed of two teams. Each team contains five players who do not know one another (see Figure 1). Matches last about 20-50 minutes. League of Legends (LoL) is currently ranked first in the world in terms of hours played for a PC game (Gaudiosi, 2012). There are about 32 million registered players worldwide (2011). The game is free to play.

Acknowledgement: We thank our interviewees as well as anonymous reviewers.


Copyright is held by the authors.
Method

We studied League of Legends through immersive ethnographic fieldwork including participant-observation, interviews offline and online (through an instant messaging tool), the collection of game logs (the game logs are available through the Tribunal System which is accessible to the player community and we selected 100 game logs that were representative in terms of types of anti-social behavior), and documents such as LoL-related forums, websites and players' blogs. In October 2011, the first author created an account on the North American Server and played about 900 matches. We conducted ten in-depth, semi-structured interviews, including interviews with six players from China and four from America. We asked them when and why they started to play LoL, what they liked and disliked about it, what forms of behavior they disliked, what they did towards the behavior they disliked, what they expected to be done to those who behaved badly, and whether they themselves had behaved in a way they considered anti-social.

Findings

In this section, we discuss flaming behavior in LOL and the Tribunal System as a method to regulate flaming as well as other anti-social behavior.

Flaming

Our findings indicate that flaming occurs most when a team is losing their game. One or more players blame others whom they believe have performed worse than could reasonably be expected. Those who are blamed accuse back. At this point, blame escalates into a flame war. Below is an excerpt from our collected chat logs. Nidalee, Ziggs, and Shen are characters in LoL. At 28 minutes, when a game might be nearing its end, Nidalee began to flame Ziggs, who talked back:

Nidalee [00:28:13]: yea stop talking ziggs
Nidalee [00:28:14]: you suck
Shen [00:28:30]: plz just surrender
Ziggs [00:28:33]: and you are pro or what?
Ziggs [00:28:34]: haha
Ziggs [00:28:35]: idiot
Nidalee [00:28:37]: 3:4
Nidalee [00:28:39]: better than 1:7
Nidalee [00:28:41]: idiot
Ziggs [00:28:48]: score says nothing
This conversation began when the team was going to lose the game, indicated from Shen's plea for surrender (players can vote to surrender if defeat seems unavoidable. But surrender is only successful when four or more out of five players in the team agree). In this conversation, Nidalee said his score is 3:4, indicating that he had killed three enemy players and died four times, which was a better score than Ziggs' 1:7. Then he blamed Ziggs for "feed," which is the act of being killed repeatedly, and thereby assisting the enemy team. In each match, there is of course always a losing team. At the moment of losing, flaming may erupt.

Flaming is sometimes started by minor, subtle triggers. Below is an excerpt in which Sion suddenly burst into a rage at the second minute of a match because he thought Fiddlesticks was positioned in the wrong place:

Sion [00:01:23]: fiddle go the fuck bot your retard
Fiddlesticks [00:01:31]: dude
Fiddlesticks [00:01:34]: watch the language
Sion [00:01:37]: dont give a fuck go bot
Fiddlesticks [00:01:38]: no excuse for that
Fiddlesticks [00:01:40]: dude
Fiddlesticks [00:01:42]: watch the language
Fiddlesticks [00:01:45]: no excuse for that
Urgot [00:01:53]: kids are here

"Bot" is short for bottom lane, one of three lanes (i.e., top lane, middle lane, and bottom lane) between two teams in a match. Usually only one player from each team is in the middle lane. Sion found Fiddlesticks in the middle lane with him, so he spoke to him in an aggressive way to force him to move from the middle lane to the bottom lane. Fiddlesticks reminded Sion to watch his language.

Players in our study said that flaming is the most detrimental form of communication against developing an effective team. One interviewee said that:

It's fine even if one player is doing very badly, but the game is unwinnable if only one player is flaming.

"It's fine" means the game is still winnable, which is most important to LOL players. But flaming upsets everyone on a team even if it happens only between two players. When players are no longer in the mood to play, they cooperate less and loss is nearly inevitable. Regarding the impact of flaming, one player wrote in a blog post that:

Flaming is perhaps the biggest reason for defeat in this game. Telling a person that he has made a mistake in a rude way, as if he doesn't know that he made it, this helps no one at all.

This player realized that flaming does not compensate for the influence of a mistake. Instead, it can only make the situation worse by making the flamed player feel bad, which "helps no one at all."

On American and Chinese servers, players often assume flamers to be people who are young, immature, and not able to control their own behavior. On American servers, flamers are sometimes called "kids." Blow is an excerpt from our chat logs:

Renekton [00:36:05]: he cant even count to 4
Sona [00:36:08]: hes so good at math
Renekton [00:36:10]: holy shit
Renekton [00:36:13]: worse than i expected
Caitlyn [00:36:16]: man u kids need to get reported for being annoying

Renekton, Sona and Caitlyn are characters in LoL. In this conversation, Renekton and Sona flamed Caitlyn, and then Caitlyn called them "kids," and said they were annoying.

In China, flamers in the game are called "pupil" in a metaphorical way. In Chinese Internet slang, pupil denotes a person who talks or behaves in an immature or irresponsible way as if they are too young to graduate from elementary schools. In LoL's Chinese official forums, some players mocked flamers by saying "Are you a pupil? Go home and do your homework instead of playing games!" Similarly, in Taiwan, flamers are believed to be junior high students or vocational school students (Sun, 2005).
The Tribunal System

Flaming, together with other anti-social behavior, cause Riot Games to implement a regulatory system in such a large player community. They knew they did not have the staff to clean up the community themselves, so they devised the Tribunal System as a way of letting the community police itself (Senior, 2011). This system combines player regulation and code regulation in one.

The Tribunal System allows a player to report other players he believes have misbehaved during a game. The player can select from a list of misbehaviors the game provides, and add additional comments (see Figure 2). However, currently reporters do not know if the reported players are actually punished. Reporting behavior is itself regulated as well. Players must select one kind of misbehavior from a list provided by Riot Games, and can only report misbehavior after a game ends.

Figure 2. Report Anti-Social Behavior.

The Tribunal System empowers players to judge reported players. Not all players have the right to judge other players. The system regulates eligibility by checking the identity of the players who log into the system. First, the players must own an account that is level 20 (the beginning level is one and maximum level is 30 and players receive "experience points" by playing games, which count toward reaching the next level). Second, the players should have never been banned before (ban is a temporary suspension from the game. It is incurred if the player has been reported to the Tribunal System and voted to be punished). If the proper conditions are met, a player can log into the Tribunal. After entering the Tribunal System, players must first agree to general guidelines that ask them to review cases in a rational and careful way. When reviewing cases, the reviewer can make a decision after 20 seconds. (Riot Games set this time constraint to prevent reviewers from reckless judging cases.) For each case, the reviewer has three options: to punish, to pardon or to skip the case.

Figure 3 shows a case in the system. A player has been reported by six players in two games. From top to bottom, the case page shows related information including game length and game type, reasons for reports, in-game chat log, and players on the same team, including the reported player. The words of the reported player are underlined in purple. At the top right and the bottom right two buttons indicate whether to punish or pardon the reported player as well as a link to skip the case. Players must make decisions alone, and cannot discuss the case with other reviewers in the system. Each case will be randomly sent to reviewers (the number of reviewers is automatically decided by the system), and if the majority votes to punish the reported player, the Tribunal System will automatically send the reported player a warning email if it’s the first time that the player is voted to be punished. From the second time, punishment is in the form of temporal suspension, which will continue to escalate if the player is punished again. Reviewers can find out the result of each case they have judged after a short period. The Tribunal System regulates reported players by suspending their accounts, and regulates reporters and reviewers by creating constraints and general guidelines for them to use these functions. This regulation is accomplished through code.
During the first year of the Tribunal System, Riot Games reported: “More than 47 million votes have been cast in the Tribunal; 51% of Tribunal cases resulted in a guilty verdict, with only 5.7% earning a permanent ban; 50% of players warned by the Tribunal just once never end up there again.” Thus the system was active and seems a serious and at least partially successful effort to control anti-social behavior in an online space.

**Discussion**

To deal with flaming, as well as other anti-social behavior, Riot Games developed the Tribunal System that aims to bring human judgment together with code to regulate behavior. The code, in the form of the Tribunal, regulates players who perpetrate anti-social behavior, players who report misbehavior, and people who review cases. Reporting may have a ripple effect beyond a single team, thus affecting the larger community. When a player violates established norms, his teammates will report him because he has ruined their gaming experience. They may also ask the other team to report the player, too. The community is made aware of the Tribunal System as a regulator. However, the code cannot work alone. It relies on the consensus of reviewers who do not know each other. By making reviewers anonymous to each other, Riot Games wants to guarantee that each reviewer's decision is not affected.
by others, and the Tribunal System is not abused. The Tribunal utilizes human judgment, but minimizes the information reviewers know about the reported player and the information reporters know about the players they have reported. In this sense, reviewers and reporters are objectified (Ekbia & Nardi, 2011) as voting elements in the Tribunal System. The Tribunal System is a sophisticated platform that brings human judgment together with regulation through code.

In his book *Grief Play Management*, Foo (2008) discussed the possibility of a player judiciary system and pointed out three challenges in implementing such a system: the selection of appropriate players to assess their peers, the ability of players to arrive at fair assessment, and the full support any such system would require of the game operator. These challenges are tackled in the Tribunal System by regulation through code along with human judgment. The Tribunal System seems effective in building a better environment as Riot Games’ statistics showed that half of the reported players no longer misbehaved again.

The Tribunal System represents a promising approach to regulating anti-social behavior. We can envision designing regulatory methods for other online venues which could leverage the participation of ordinary people in regulation along with code.

**Conclusion**

Flaming has been problematic since the World Wide Web made internet services widely available. The Tribunal System tries to bring human judgment into code regulation and is reported to be performing well. But it is complex to combine human judgment and code regulation. There are still many challenging issues in designing such systems, such as deciding how much information should be made available to human judgment. There also exist questions such as how people use the system and how this information could be utilized in regulating anti-social behavior in other online venues. In future research, we will try to answer these questions to help design better regulatory systems.

**References**


Cooperative, Dynamic Twitter Parsing and Visualization for Dark Network Analysis

Patrick M. Dudas
School of Information Science
University of Pittsburgh
pmd18@pitt.edu

Abstract

Developing a network based on Twitter data for social network analysis (SNA) is a common task in most academic domains. The need for real-time analysis is not as prevalent due to the fact that researchers are interested in the analysis of Twitter information after a major event or for an overall statistical or sociological study of general Twitter users. Dark network analysis is a specific field that focuses on criminal, terrorist, or people of interest networks in which evaluating information quickly and making decisions from this information is crucial. We propose a platform and visualization called Dynamic Twitter Network Analysis (DTNA) that incorporates real-time information from Twitter, its subsequent network topology, geographical placement of geotagged tweets on a Google Map, and storage for long-term analysis. The platform provides a SNA visualization that allows the user to interpret and change the search criteria quickly based on visual aesthetic properties built from key dark network utilities with a user interface that can be dynamic, up-to-date for time critical decisions and geographic specific.

Keywords: dark networks, visualization, social network analysis, user-design

Introduction

When looking at patterns in large datasets, a popular choice is the utilization of Twitter. It provides user-generated content, or micro-blogging, in a real-time environment as a social media outlet that is propagated from one individual to another in terms of text, links, pictures, or videos. Projects for this type of networked data is usually data-mined after the fact, because either researchers are interested in a certain past event (Lotan et al., 2011; Yardi & Boyd, 2010) or topics of interest (Tumasjan, Sprenger, Sandner, & Welpe, 2010). There may also be a need for a very large dataset and backtracking provides the best means of collecting ample amounts of information versus real-time information (Bruns, 2011; Cha, Haddadi, Benevenuto, & Gummadi, 2010; Kwak, Lee, Park, & Moon, 2010).

Dark networks (Raab & Milward, 2003) or terrorist informatics (Cheong & Lee, 2011) provide their own unique problem set because immediate reaction is critical to the success of authorities looking for terrorist or criminal activity. We propose an architecture and visualization to enhance user decision making using various saliences (Lurie & Mason, 2007) from key utilities designed for dark networks (Roberts & Everton, 2011). We also connect this information processed in Twitter and map geotagged tweets to a Google Map, combining both network topology and spatial information. We built a Twitter feed that allows the user to click links embedded in the tweet directly so the user can see if a particular website, social feed, or video is causing the propagation of information or other sources they should consider viewing. An additional visualization allows the user the ability to see temporal pattern changes in the network. All of this tailored to be simplistic and individualized allowing the user to input a location and either a username, hashtag, or keyword to pull information and begin visualizing.

The novelty of this approach is that we built an interface that can process, visualize, and store real-time Twitter data in network visualization and incorporate various salient features to capture the most prudent information. This platform is also novel in that as a website, the user can log in and work on the

Copyright is held by the author.
same dataset as another user by selecting that project. Multiple people can view multiple locations and multiple terms, but for a common goal in capturing and analyzing all information obtained in that project. We provide two mechanisms of visualizing: 1) Real-time network topology, and 2) Long term network analysis.

The project was built and designed for use-cases surrounding the Syria uprising, a topic being analyzed at the CORE Labs at the Naval Postgraduate School (Monterey, CA). Since its development, it has been showcased to a variety of different users and groups in or associated with various government agencies. Provided in the paper is also a high-level view of the data from locations of interests.

Background

More recently, a variety of different types of real-time architectures have emerged in handling and interpreting Twitter data. Murthy suggested a means of employing real-time data into a data repository/data gathering application in terms of research related to cancer (Murthy, Gross, & Oliveira, 2011). This novel approach provided a framework, but lacked the ability to visualize or create a user interface for sense-making when it came to harvesting data.

A real-time data approach to terrorism informatics was developed by Cheong (Cheong & Lee, 2011), who provided a detailed look at parsing and filtering the data from Twitter. They provided some markups on how they would visualize this information for users, including: 1) WEKA timeline analysis (Hall et al., 2009); 2) a Google Map with geotagged tweets; and 3) a self-organizing map for unsupervised clustering (Kohonen, 1984). The timeline analysis is a great mechanism for visualizing terrorism or dark networks, but lacks the ability to network analyze in real-time. Also, the Google Map approach was strictly a markup. We provide this functionality along with the ability to select the geotagged position to view the tweet involved with this geospatial information.

SocialAction (Perer & Shneiderman, 2008) extends this model by interweaving the statistical properties in the visualization itself, imploiring a higher rate of success of exploratory SNA. They even extended this expert knowledge to analysis of terrorist connection. The restriction with this approach is that the only information passed to the user is network topology and the salience changes based on the embedded statistics of that network. It lacked real-time data which is crucial to making decisions quickly and efficiently.

Visualization

Real-time

We developed a web application that allows an analyst the ability to: 1) collect information on Twitter by hashtag, username, or keyword; 2) visualize this information or network in real-time; and 3) change initial search criteria based on the visualization and applied social network analysis properties. The process starts by having the user log into the website and selecting either an existing dataset or creating a new one. Once a project is selected or created, the user then inputs city, state, country or zip code and a search term. For the search term, s/he can either select hashtag, username, or keyword to search. The web-application then utilizes Google Maps API (Google, 2005) to geocode the address given. Lastly, the user can select the type of network s/he is interested in collecting and viewing, which includes: 1) username -> username 2) username -> hashtag 3) hashtag -> hashtag. A mock example is given below.

Example:

James tweets: “Just saw @Dannie at the #Pittsburgh #Library”

Username -> Username: James -> Dannie
Username -> Hashtag: James -> Pittsburgh and James -> Library
Hashtag -> Hashtag: Pittsburgh -> Library

With the geocode, search term and network type specified, the application creates a global search query, which is communicated to Twitter’s Search API while the page is active. The entire architecture is listed in Figure 1.
At this point the user is provided the D3.js (Bostock, Ogievetsky, & Heer, 2011) force-directed network visualization, Google Map, and Twitter feed, all of which continually updates as long as the website remains active. The Twitter data is saved to a MySQL database, including the timestamps of tweets. The network itself has built-in aesthetic saliencies that suggests to other users, hashtags, or keywords that may provide better insight into the original search topic.

**Saliency and Network.** The graph for the tweets is created every time a query is requested from Twitter. The network, when visualized, is limited to only 200 nodes to limit the amount of attention needed by the user. The nodes themselves are selectable and will augment the original search query to include this hashtag, username, or keyword. This is a measure to help limit the amount of noise in a given graph. A clustering algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) is applied to the network and groupings are assigned and represented by color assignment. Based on log-based degree centrality, the size of the node reflects the number of connection each username or hashtag accompanies. Considering Twitter is a directed network, edges reflect the direction of the message. Figure 2 highlights both the groupings and centrality, with a bias towards increased inner-cluster strength versus out of cluster nodes. The importance of these saliency changes is to highlight nodes of interest or key players, which is important in dark networks (Everton, 2012).
Geographical Display and Twitter Wall. We appended a Google Map that shows geotagged tweets with the tweets themselves (Figure 3). This allows the user the ability to view the tweet and from where it originated. We created a text GUI that displays the tweets (Figure 4) along with the URLs, hashtags, and users. The URLs are selectable and will show the website in a new window. This will allows the user to see if a news article, YouTube video, blog, social networking post, etc. is causing an increased number of tweets.

Figure 2. Network Visualization

Figure 3. Google Map with tweet

Figure 4. The Twitter Feed (Including: username, the tweet, and timestamp)
Long Term

We provide a mechanism to group various datasets together for a single visualization. This can then be outputted as a .gexf, .net, or .dl network type. To visualize this closer to real-time, we developed a stream in a social network analysis application called Gephi (Bastian, Heymann, & Jacomy, 2009) and uses Gephi Streaming Network add-on to view the network as it evolves. Figure 5 below contains only 5 minutes of data collection and using the username -> hashtag connection. It was a grouping of both Syria and FSA search terms in Syria.

![Gephi Visualization](image)

**Figure 5. Gephi Visualization**

We also developed our own means of visualizing entire projects called TweetViewer. This web application is to be used with DTNA and will display the completed network. Figure 6 shows the network in TweetViewer. TweetViewer for DTNA promotes the temporal data associated with the network by providing a timeline to allow a user the ability to specify at which point in time they are interested in viewing the network or the network’s evolution over time. As DTNA’s network visualization, edges reflect the direction of the message. Additionally, we differentiated the nodes that originated from the tweet by augmenting and animating the nodes’ borders as time progresses. Figure 7 shows TweetViewer as it progresses through the timeline. Lastly, users can search based on the tweet to showcase nodes that include certain text. Figure 8 shows a network built using the hashtag “#FSA” in Homs, Syria and is highlighted by tweets that include the word “Syria.”
Figure 6. TweetViewer visualization

Figure 7. Network in TweetViewer using timeline
Current Usage and Initial Findings

This project was only recently developed, but has already been employed by 30 users from a variety of military groups or organizations, ranging from civilian use to higher ranking officers including majors, first lieutenants, captains, and lieutenant colonels. This usage has accounted for 182 queries and has pulled roughly 86,000+ tweets as of publication. The majority of the tweets have centered on search areas (geographically) in Syria and the Philippines. Only 18% of the tweets pulled from these areas are geotagged. The most common search term is the hashtag “#FSA” for Free Syrian Army. There were roughly 8,000+ URLs in each tweet, which suggest a high influence on external websites driving the tweets.

Future Development

There are three avenues of research that we are currently interested in pursuing: 1) visualizing sentiment in Twitter posts, 2) determining the value of networks created using either username -> username, username -> hashtag, or hashtag -> hashtag connections, and 3) visualizing temporal data using more specific temporal dependent cluster algorithms and network centrality.

Sentiment Analysis Visualization

We are interested in not only the network that develops from this application, but also the sentiment for keywords. The way we are looking to address this is by implementing a sentiment lexicon (Nielsen, 2011), which includes word lists that are given a score from 5 to -5, based on their positive or negative connotations, respectively. This could be used to evaluate social movements or key terms and how over time these connections are influenced via sentiment lexicon over time. This does not have to be limited to strictly a global overview. It can be topic-based as well, allowing for searching of words or parts of words and connecting their positive or negative connections. The sentiment lexicon would be fixed around a circle, and the keywords would be placed in the center (Figure 9). Based on the number of occurrences of a key word or phrase in conjunction with a sentiment word would determine the nodes.
polarity towards the positive or negative side of the circle. Below are two instances of this analysis at two different timestamps (Figure 10).

Figure 9. A markup of the sentiment visualization

Figure 10. A markup of the sentiment visualization as the nodes shift from one side of the circle to the other (from negative to positive)
Value of Twitter Hashtag Networks

As of this paper, there has been little to no work done on the quality or even the characteristics of a network built on a variety of different measures, such as: 1) username -> hashtag, or 2) hashtag -> hashtag. Username -> username is quite common, but username -> hashtag and hashtag -> hashtag may yield new insights into social structures. This would go beyond a social semantic ontology in that the user specifies these keywords (hashtags), which connects both a personal meaning by the person tweeting the information and by a global meaning in the concatenation of keywords (hashtags), which could be “trending” in the Twitter realm.

Visualizing Temporal Data

We are looking to implement some characteristics that are becoming popularized in evolving networks into our TweetViewer visualization. There is a need to not only show a network over time but to provide a visualization that allows for prediction. Macskassy presents multiple factions of evolution that could be applied to a visualization (Macskassy, 2012). This includes: 1) SNA metrics, such as Bonacich Centrality (Bonacich, 1987), with temporal parameters, and 2) a machine learning approach to predict and forecast. These measures take into account that communication evolves over time and should not be judged simply as a single, static, aggregated snapshot.

Conclusion

In this text we outline a cooperative, dynamic, and interactive visualization tool to collect and navigate through Twitter tweets. This includes a dynamic network visualization, a tweet wall post, and map with geotagged tweets. Additionally, we create a separate but homogeneous visualization that allows for exploration using timestamps and text from the tweets themselves. This application has been implemented into military outlets and has received high praise for its ability to data mine. We believe this tool has a plethora of interesting usages in dark networks and provides real-time information to people who most need this type of data.

References


Abstract

In Library and Information Science our emphasis is typically on facilitating reuse of documents. There are many motivations behind this emphasis, but efficiency and time savings are common motivations. However, participants in an on-going study of knowledge communication practices have a strong tendency to re-create existing documents from scratch, despite the fact that sometimes multiple documents that serve the same purpose already exist in the organization’s archive. While at first this behavior seems wasteful, examination of their practices indicate that participants often have good reasons for re-creating documents. These include (a) empowerment, (b) learning, (c) personalization, (d) customization, and (e) simplification.

Keywords: information behavior, knowledge management, social and community informatics

Introduction

Why use an existing document when you can create one from scratch? While this statement seems at first glance purely rhetorical, in a study of a high turnover voluntary organization it was surprising how often documents were created from scratch, despite the fact that similar documents intended to serve the same purpose already existed. At first glance, this fact is puzzling, and seems to indicate something is going seriously awry. Perhaps the organization is disorganized and cannot keep track of its documents. Perhaps it is extremely inefficient, constantly making work for itself. However, when the practices surrounding the re-creation of existing documents are examined, it turns out that the re-creation actually serves some important organizational purposes, and that naT, ve efforts to “streamline” the organization by reducing the amount of document re-creation might actually cause more problems than fewer.

This analysis goes beyond classic research questions in knowledge management (KM) (e.g., Alavi and Leidner (2001)) to look at what other organizational functions seemingly wasteful documentary practices might have. The goal of this research is to identify possible candidates for generalizable unintended consequences of introducing KM solutions (whether social, technical, or sociotechnical) before the solutions are introduced (Schultze & Boland Jr., 2000). In particular, it does not limit itself to looking at unexpected KM consequences, but rather looks for possible unexpected organizational consequences.

A Bumble-Bee Organization as a Research Site

This note reports some preliminary findings of a study of knowledge communication practices with a focus on the functional role that documents play. Three major methods were employed: participant observation, interviews, and document analysis. There were 46 participants whose activities were actively studied in over 200 observation sessions and interviews over the course of a year (October 2011-October 2012). At least 500 documents were analyzed in varying levels of detail as a part of the research. In my participant observation I took an active role in the organization, volunteering my time for the organization.
in order to participate in it so that I would fully understand how it functions, rather than serving as a passive observer. My data collection and processing involved extensive writing of field notes. Documents were analyzed for their functional features in facilitating knowledge communication based on how they were actively used by participants. The research focused primarily on social interaction around documents both face-to-face and via electronic communications (primarily email, but also Google Docs and other social software). Interviews were scheduled with participants when either they did something interesting in order to understand why, or to obtain their perspective on observations that I was making. Most of the interviews were very informal, as I found that turning on the tape recorder prevented people from speaking freely about why they did what they did.

The Researched Graduate Student Union (RGSU), a union for graduate student employees at a large Midwestern university, was selected as the site for this research because it had several important features. First, it is an organization with high turnover: on average active members stay active for about a year; the leadership changes almost completely from year to year. The bargaining cycle which determines the lifecycle of the union is typically 3 years, meaning that few people are still active in the organization from one bargaining cycle to the next. Paid staff (who also participated in the study) stay a bit longer, but only one staff person has been at the organization for over 4 years in its entire history, and his role is that of a part-time administrative assistant. Many members never meet him.

Second, the organization has a strong commitment to preserving its knowledge: it has a large archive of documents that date from the very beginnings of the organization; a significant amount of effort is made to save documents and keep them accessible; in most meetings careful minutes are taken by staff and preserved to keep a record of the actions and decisions of the organization.

Third, while the organization seems to be on the brink of falling apart as knowledge of core skills and organizational memory are constantly being lost, it still manages to be extremely successful both in performing its core functions and in using the same methods to successfully pursue its goals. We might call it a “bumble-bee organization”: it seems like it should not be able to fly, yet it flies extremely well.

The fact that documents are re-created so frequently in the RGSU is particularly interesting case of this bumble-bee nature because there are a number of pressures that would seem to encourage document re-use over document re-creation. First, the two most valuable quantities in the RGSU are labor and time. In many ways they are seen as the currency of the organization by most experienced active members. The RGSU is a large organization with about 2,700 members. While the RGSU hires 3 permanent staff, the bulk of the work of the organization is performed by active members, all unpaid volunteers. New people are constantly becoming active while more experienced people drop out—typically there are about 40-70 active members. Nearly all of them are also working and all of them are students. Furthermore, one of the core values of the RGSU is the value of family and friends, so most active members have significant social and/or familial commitments as well. Thus, the amount of work a particular member puts in (labor), and the amount of time they have to do it, are both a measure of commitment and a scarce resource, and active members are constantly struggling to balance all of their commitments. Therefore, any time spent doing something that is unnecessary is seen as wasteful and active members will quickly become frustrated when a task is perceived as such.

Second, every participant in the organization is an expert at using and managing documents simply due to the requirements of being a graduate student. Evidence of this includes the large number of documents that are created and used, the volume of intra-organizational email messages (over 100/day), and the sophistication with which the documents are structured and organized. Members are keenly aware both that the re-creation of documents occurs, and of how unnecessary it seems. The general sentiment when this is discussed in the organization is one of embarrassment: most participants feel that they and the organization ought to be better skilled at managing documents.

The Hidden Benefits of Document Re-Creation

Why is document re-creation so frequent? Analyzing the practices of document creation and the functional roles they play in the organization reveals five purposes that document re-creation serves.

---

1 Thanks to Dr. Michael Twidale for this name.
Document Re-Creation as Empowerment

Because of the time and work pressures outlined above, participation in the RGSU must be rewarding for active members or they would not volunteer so much of their time. There are many factors (community belonging, community of shared values, etc.), however one of the most important rewards is a sense of ownership, pride-in-work, and respect. These feelings are empowering, especially given the contrast to graduate school where many active members feel subject to the whims of their advisor, their department, or the university as a whole.

The opportunity to create a document that reflects hard earned knowledge, the process of creating an artifact for the organization, the trust of fellow active members embodied in their encouragement to create the document, and the respect of fellow active members that a well created document engenders, all are extremely motivating, and create a sense of empowerment, belonging, and shared participation in a communal effort.

For example, one common way in which a person moves from being more on the legitimate periphery (Lave & Wenger, 1991) to being more centrally active is by performing the labor of creating a needed document. This can be a flyer, a hand-out, a guide, or even updating an existing document that needs regular edits. It is not uncommon for more experienced members to encourage newer members when the latter are excited about a task of creating a document, even when the experienced members know it probably need not be created entirely from scratch. What they do not want to do is prevent the newer member from having an opportunity to become more fully involved with the organization; a little bit of redundant labor is worth the trade-off of having a newly empowered active member. Community growth and individual respect is always more important than minor gains in efficiency.

We found that RGSU staff are much less likely to re-create documents than active members. There are many reasons for this. Regarding empowerment, staff are empowered by being central to the RGSU’s daily functioning, and need not create documents to remain there. They still do create documents, but it is typically on an as-needed basis. Only rarely will they find a need for a particular document that they are passionate about creating.

Document Re-Creation as Learning by Teaching

There is evidence that teaching a topic to someone else greatly facilitates learning (Grzega, 2006). Similar to teaching a class, the activity of creating a document can provide an opportunity for a participant to be reflective on their own practice, and crystalize their experience in such a way that it can be of use to someone else.

For example, a long-time active member (M1) re-created a document: a how-to guide for running department meetings (DM1). This replaced an older version created a year prior (DM2) by a different member (M2). There is at least one more version of this document (DM3) created by a third member (M3) in the RGSU archives. Each document has been developed completely independently. M1 was an active member when M2 created DM2, and she was aware of its existence. While I currently do not know who M3 is, I do know that the creation of DM3 predated both M1 and M2 in the organization, and that when M2 created DM2, she was unaware of the existence of DM3.

One of the most important reasons that both M1 and M2 created their respective documents was to encode their learning into a document. M2 in particular spent weeks on her document as part of a process of reflection on her own experience trying to run successful department meetings in different departments that had a very different nature, and to communicate that experience and make it accessible to other members. M1 has a very different kind of experience. She has been very active and successful in organizing her home department (that has over 100 members), but she had heard that DM2 was long (5 pages) and challenging for new stewards to use. Given her experience with new stewards being reluctant to take on running department meetings, she took her experience in the Stewards’ Assembly (supplemented with feedback from staff and other stewards) and focused on creating a 1 page document whose express purpose was to “demonstrate how little labor it takes to run a department meeting.”

Document Re-Creation as Personalization

The personalization of a document by re-creating it is a powerful way of having ownership of a part of the RGSU. This phenomenon is often coupled with one or more of the other benefits. Thus, a part of empowerment is having a personal touch on the document that is created. And a part of teaching is
putting the emphasis on the material that you feel is most important, as derived from the lessons you have learned through experience. There are many layers to personalization, and it is worth exploring them a bit.

The lightest layer is a personal style in the document’s creation. It can take many forms, from subtle in-jokes embedded in the text or structure, to not so subtle personalized vocabulary. Whatever the form may take, this kind of personalization is typically little more than a flourish that identifies one or more of the authors as such, and may provide a bit of entertainment to active members, but not more. The fundamentally communal nature of the RGSU means that people who feel the need to control details or leave their personal mark as a means of control typically do not last long, because for the RGSU to work effectively people must feel an equal, shared sense of control without any one person taking primacy. Even the role of the co-presidents is more that of mediator, and their leadership is most effective when they amplify the thoughts of other members rather than try to explicitly push their own agenda. The personal touch is most effective when it is simply a signature indicating a public performance of labor, and is insufficient alone to motivate a person to re-create a document.

The deepest layer of personalization is when a core understanding learned through experience is translated into document form (e.g., M2). This kind of personalization is often expressed over days, weeks, or even months crafting a document. Many documents in the RGSU have such an origin, including the previous version of the guide for running department meetings (referenced above), the Stewards’ Handbook, various strike reflections, and several strategic plans. Interestingly, these documents are often some of the less used documents; whether because of their length, their density, or for some other reason is not yet clear. This kind of deep personalization is sufficient motivation to re-create a document. However, it can also motivate a thorough editing of a document instead of a re-creation, thus other factors are also at play.

Levels of document personalization can of course range between these different layers.

Document Re-creation as Customization

One of the most important purposes that document re-creation serves in the RGSU is customization for current circumstance. The RGSU is in a constantly changing environment, where the issues it must address and the circumstances it must respond to are almost completely out of its control. It is vital that the RGSU remain able to adjust to unexpected changes or crises.

Documents are re-created within the particular context the organization finds itself in at the moment. Thus, they are created to teach people what is not currently common knowledge, or to be structured in such a way that they are accessible to the people who are currently a part of the organization.

Continuing the example above, M2 was created when membership was low, bargaining was not yet in progress, and it was difficult to convince people to run department meetings. It was created from the perspective of a member who had tried many different strategies in her home department and still was struggling to find a sustainable solution. She had also worked with several other departments to help them run a department meeting. Consequently, her document was several pages long, emphasized the many different options that existed for having successful department meetings, and left a lot of room for reflection, innovation, and exploration.

In contrast, the newer document for running department meetings was intended to be a quick, simple, bare-bones guide. It will not necessarily work for all departments, but the checklist on a single sheet of paper is easy to follow, meaning that new and busy stewards in the middle of ramping up for bargaining have a simple recipe to follow that they can then build upon. This document thus reflects the current state of frantic activity for which the other document is much less appropriate.

Document Re-creation as Simplification

All of the preceding benefits of document re-creation can serve as lesser or greater motivation for the activity. However, none of them necessarily lead to document re-creation. One of the major factors that tip participants towards re-creating a document is the well known factor of the ease of old document retrieval from the archive. Rather than spend much time on an oft-explored issue (e.g., Berlin, Jeffries, O’Day, Paepcke, and Wharton (1993) and their exploration of group memory software), I will simply review a few of the most relevant factors for the RGSU archive.

First, the RGSU archive is scattered over several different media. There is an extensive paper filing system, there is a file system currently split between two office computers, there is an online dropbox with files, and there is a massive collection of Google Docs. Not everybody has easy access to each, and
this is further complicated by the fact that both Dropbox and Google Docs have displayed problems where the reliability of their data is, at times, questionable. Therefore, there is the issue, can the document be found? Where should I look for it? Who put it where and why? Etc.

This is further complicated by the fact that, even though the filing systems have each been organized on several different occasions by people aiming to make them more accessible, it is still not clear at first glance what the actual organizational theme is by which the documents are arranged. This of course means that any newcomer to the filing system must try to figure it out first, which takes a non-trivial amount of time. Thus, in many cases, re-creating a document from scratch, especially when it either is on a topic that is well known, or when it provides one of the other benefits, can actually be simpler if not quicker than trying to find a document that may or may not (still) exist.

Conclusion

Despite the fact that re-creating documents looks, at first glance, like a waste of time and effort, there are actually good reasons for re-creating certain documents. These reasons include (a) empowerment, (b) learning, (c) personalization, (d) customization, and (e) simplification. Empowerment is important for engaging peripheral members into the active core group of members who provide some level of stability to counteract the high levels of turnover in the organization. Learning is important because it is one of the means by which members develop their knowledge of the organization more fully, and create artifacts that can be used to teach it to others. Personalization helps motivate some members to do work for the RGSU. Customization allows the created documents to be relevant to the current state of the RGSU, including documents that at first glance might be thought to have a single “best” generic structure. However, while simplification may motivate any of the other four reasons for document re-creation, it also suggests that some inefficiencies exist in the documentary archive of the RGSU which bear further examination.

Members of the RGSU are concerned with the amount of document re-creation that exists in the organization. In future work I hope to examine when document re-creation is appropriate, and when it is unnecessary. Part of this will involve getting some sense of how frequently it happens for each of the above reasons, part of this will involve more closely examining the frequent cases of document updating and reuse, and part of it will be examining the means participants have for effective document discovery.

References


Grzega, J. (2006, September). Developing more than just linguistic competence. the model ldl for teaching foreign languages with a note on basic global english. Humanizing Language Teaching, 8(5).


Hanna Maurin Söderholm  Monica Lassi  Gustaf Nelhans  Jan Nolin

hanna.maurin@hb.se  monica.lassi@hb.se  gustaf.nelhans@hb.se  jan.nolin@hb.se

Swedish School of Library and Information Science, University of Borås

Abstract
This study investigates self-reported practices and values of Internet-based cultural consumption in relation to illegal file sharing in Sweden during 2007-2012. It is based on more than 1,000 posts in an Internet discussion forum and focuses on expressed senses of guilt, whether file sharing is right or wrong, buying media, and dimensions of stakeholders and media genres. This note presents work-in-progress results of initial quantitative analysis, suggesting that no changes in feelings of guilt were detected during the time period, despite legislative changes. More posts report to now buy media, while there at the same time is an increase in posts expressing that file sharing, despite its current illegal status, is a right thing to do. A deeper qualitative analysis is needed to further understand the complexity of changes in file sharers’ justifications for what content to acquire through illegal file sharing, what they choose to pay for – and why.

Keywords: Internet-based cultural consumption, file sharing, guilt-feeling, right–or–wrong, Flashback.org

Introduction

In the context of Internet-based Cultural Consumption (IBCC) discussed by Nolin (2010) more and more people are using the Internet as an arena for their cultural consumption. IBCC entails different genres such as music, movies, games and e-books, and legal and illegal file sharing are established means for acquiring this content. In the history of IBCC file sharing some key points can be identified with respect to: technological development and applications such as the MP3 codec in 1987, Napster, Kazaa, and the BitTorrent technology; and, legislative actions taken by government agencies, such as The Digital Millennium Copyright Act (1998). One of the main problems within the legislation drive during the 1990s’ was the difficulty of policing millions of users within each national context. A global trend toward solving this problem by decentralizing file sharing legislation, led to the EU directive on the enforcement of intellectual property rights (IPRED). IPRED came into force in May 2004 and directed member states to comply with it within two years.

Sweden, a country characterized by early adoption of Internet technology in general and as perhaps the most prolific European country in file sharing technology development and use, was a late adaptor. File sharing was widespread and had been steadily increasing during the 2000s’ and the pioneering file sharing site Piratebay.org was hosted here, attracting users from all over the world. It could be argued that in Sweden, the threshold towards sharing material (both copyrighted and own) was low, due to the fact that there already were legislated procedures allowing private copying material and sharing it in closed groups between friends and family, i.e. ‘fair use’. Thus, the IPRED implementation was highly controversial and led to several internal rifts within leading political parties. Right before the law, the debate among government agencies and copyright owners was intense and far from obvious when it came to outcome and opinions.

In April 2009, IPRED came into force. Overnight, file sharing practices among millions of Swedes went from being a (relatively) socially accepted way of acquiring and consuming culture, to be re-confirmed as an illegal act. This shift happened without any visible external changes such as changes in...
technology or in what content that was possible to download. In other words, nothing really changed, except for the introduction of a law that re-confirmed file sharing as a criminal offense. How did Swedish file sharers react to the implementation of this legislation? Overall, studies suggest that the IPRED law led to some temporary changes among the general public, but that the activities now are back to the same levels as before the law. According to a study by Larsson and Svensson (2012) file sharing behaviour had, six months after IPRED was implemented, somewhat decreased while social norms seemed to be unaffected. With few exceptions, Larsson and Svensson (2012) being one, these changes have been measured through static measurements of data traffic and discussed in dichotomized terms of piracy, internet warriors, and free information on one hand and intellectual property rights and economic interests on the other.

The purpose of this work is to explore self-reported behaviors and motivations among illegal file sharers during a five year time period of increased juridical legislation and societal pressure in Sweden. This paper presents the initial results of a quantitative analysis of a discussion topic at the Internet forum Flashback.org devoted to feelings of guilt in relation to file sharing.

**Case Description: Flashback.org**

The Swedish discussion forum Flashback.org is one of North Europe’s oldest and most well known Swedish Internet forums, although often regarded with a lot of scepticism by media and the general public. According to the official mission statement Flashback.org is politically and religiously independent and its purpose is to protect and defend free speech. By international comparison it resembles the 4chan bulletin board in how it accepts and attracts a wide range of discussion topics including, but not limited to, illegal activities and phenomena. As of September 2012, Flashback.org had 685.179 registered members and 37.168.504 posts. The most active discussions are dealing with Computers and video games; Relationships; National politics, Integration and Immigration; Music; and Crime. These discussion categories consist of 0.8 to 1.3 million posts each, distributed over a large number of topics.

We chose to investigate a topic from this particular forum for several reasons: 1) it is one of Sweden’s most active arenas for discussing file sharing, as the File-sharing category alone consists of 23.000+ topics including approximately 450.000 posts; 2) the forum content dates back to before year 2000, enabling analysis of possible changes in discussions and perceptions over time. Furthermore, the specific topic we chose provides us with the unique opportunity to access perspectives from active file sharers in Sweden during a time when significant political and legislative changes occurred.

The person who started this particular topic directly addresses active file sharers, asking whether they ever felt guilty because of their activities. Thus, the topic title translates roughly to “You guys who download stuff illegally, don’t you ever feel guilty?” (In Swedish: “Ni som laddar ner olagligt, får ni inte dåligt samvete någongång?”)

**Methods: Study Design, Data Collection and Analysis**

We conducted a basic quantitative content analysis of the aforementioned topic in the Flashback forum. The topic thread was created in 2007 and exhibits three distinctive “burst” of activity, in 2007, 2010 and 2012. With the exclusion of three posts during one day in 2011, this makes up a natural division for comparison between groups over time.

**Table 1. Discussion activity distribution 2007-2012**

<table>
<thead>
<tr>
<th>Year</th>
<th>2007</th>
<th>2010</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12th-May 15th</td>
<td>1st-Dec 26th</td>
<td>8th-Aug 6th</td>
</tr>
<tr>
<td>Discussion active between</td>
<td>March 12th-May 15th</td>
<td>Nov 1st-Dec 26th</td>
<td>March 8th-Aug 6th</td>
</tr>
<tr>
<td>Number of posts included in this analysis*</td>
<td>358</td>
<td>401</td>
<td>318</td>
</tr>
</tbody>
</table>

*Empty posts, spam or nonsense (e.g. “Hey there!!”) were omitted from this analysis (n=12).

The discussion consisted of 1089 posts (in Swedish) by 796 unique anonymous authors (unique user names). We assume that the authors of this publicly available discussion are Swedish and that they are engaging in illegal file-sharing activities, because they respond in Swedish to the question posed in
the topic title which addresses illegal file sharers. Flashback.org does not provide any member demographics and there is a limited amount of information about member profiles available for non-members.

Activity wise, while the majority of the posts were unique, the 2007 period had somewhat more interactive discussion and reoccurring posters than the other two. However, there were few posters moving between the time periods, only 12 (1.5%) of the authors posted in more than one of the three time periods. Content wise, the posts could be anywhere between a simple yes or no as answer to the question posed in the topic title, and elaborate reasoning and rationales for a certain stance, spanning several pages.

We developed a coding scheme following Neuendorf’s (2002) suggested method, including several researchers in all steps of the process. We recorded e.g. whether the posts expressed feeling guilty or not in relation to file sharing, whether file sharing is considered right or wrong, buying of media, any mentions of media genres or stakeholders, and, if any of the posts expressed change of mind or behaviour at some point. The posts were manually coded into SPSS and analysed by two researchers. Data was analyzed through frequency calculations of occurrences of the items described above, first within and then across the time periods.

Results

Is File Sharing Considered Right or Wrong?

A small number (n=127, 12%) of the posts explicitly stated whether they thought file sharing was right or wrong. When contrasted across time periods, changes were detected in response distribution.

![Graph showing distribution of posts expressing file sharing to be right or wrong](Figure 1)

The responses stating that file sharing was right doubled from approximately one-fifth to two fifths. Correspondingly, the share of respondents indicating that it was wrong dropped from two-thirds to approximately the same share as their opposites (two fifths). However, there seem to be a slightly higher proportion of posts reporting “both right and wrong”. Taking into account the respondents answers that indicate both yes and no the affirmative responses rise from less than a third to more than half of the respondent between the three time periods. It should, however, be noted that the number of posts explicitly mentioned this aspect was very low in each of the time periods.

Changes in Behavior

Overall, a small portion of the total posts (n=40, 3.7%) reported changes in behavior or opinion in relation to their file sharing practice. As illustrated below, there were more posts reporting this in 2010 than during the other time periods.
After IPRED was introduced in April 2009, there were reports of decreased Internet data traffic. The first indications by general media indicated that data traffic in Sweden dropped with 30–50% within the first 24 hours after the law was in force (Lewan, 2009). These data were almost immediately criticized and the next day Pirate Bay reported that the traffic on their site dropped by a mere 3.5% (Aftonbladet, 2009). The results found in this study correspond with these studies, in suggesting that the law had a temporary scare effect.

**Buying Media**

Overall, a total number of 298 posts (~28%) explicitly stated to be paying for/not paying for cultural consumption, in the sense of monetary purchases of media.
Allegedly, the respondents in this study are involved in illegal file sharing (since that was the stated at the outset in the initial question), yet the affirming share of posts who states to be paying for some of their culture consumption actually increases from 74.3% to 94% during the time of the study. However, it is important to be careful when interpreting these numbers, as well as illustrating a change induced by the IPRED law they might simply mirror the fact that there is more content available through legal services today, such as Spotify, iTunes, Netflix etc.

Genres and stakeholders

The majority of the posts mentioned one or several media genres and/or different stakeholders. The notion of stakeholders include reasons, rationales and justifications for file sharing in relation to “personal” motives such as financial or ideological, “business” as in relation to the media industry, and “artist” as in the performer, artist or author of a particular piece.

Table 2. Mentions of stakeholders per media genre

<table>
<thead>
<tr>
<th>Genre</th>
<th>Stakeholder</th>
<th>2007 Period 1 (n=420)</th>
<th>2010 Period 2 (n=640)</th>
<th>2012 Period 3 (n=460)</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>Personal</td>
<td>0,7%</td>
<td>3,3%</td>
<td>3,9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Business</td>
<td>0,7%</td>
<td>0,6%</td>
<td>2,2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Artist</td>
<td>0,0%</td>
<td>0,5%</td>
<td>0,7%</td>
<td></td>
</tr>
<tr>
<td>Games</td>
<td>Personal</td>
<td>5,0%</td>
<td>7,3%</td>
<td>7,0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Business</td>
<td>1,7%</td>
<td>3,6%</td>
<td>2,8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Artist</td>
<td>1,9%</td>
<td>2,3%</td>
<td>3,0%</td>
<td></td>
</tr>
<tr>
<td>Music</td>
<td>Personal</td>
<td>14,3%</td>
<td>13,4%</td>
<td>16,1%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Business</td>
<td>16,0%</td>
<td>9,2%</td>
<td>8,0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Artist</td>
<td>17,4%</td>
<td>11,1%</td>
<td>12,2%</td>
<td></td>
</tr>
<tr>
<td>Movies</td>
<td>Personal</td>
<td>9,3%</td>
<td>14,1%</td>
<td>15,4%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Business</td>
<td>6,4%</td>
<td>8,1%</td>
<td>8,5%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Artist</td>
<td>7,1%</td>
<td>6,7%</td>
<td>8,5%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>19,5%</td>
<td>19,7%</td>
<td>11,7%</td>
<td></td>
</tr>
</tbody>
</table>

For example, the posts where the genre “Movies” and stakeholder “Personal” are mentioned increased from 9.3% to 15.4% during the time period. This might be interpreted as file sharing and the media genre “Movies” is increasingly discussed in relation to personal motivations and justifications and to a greater extent than aspects related to the movie industry or individual actors.

Feeling guilty

In all, 603 posts (56%) explicitly responded to the question posed in the topic title: “You guys who download stuff illegally, don’t you ever feel guilty?” The large majority of these reported not having any feelings of guilt. However, when analyzing data across the time periods, we detected small nuances.
In 2010, there was a small dip in the responses of feeling guilty. In addition, there were a slightly higher proportion of posts stating a more nuanced response, as in “both yes and no”, especially compared to 2007 – where on the other hand a larger portion of posts reported indifference to the question. This could indicate a degree of ambivalence to the practice, but at the same time also that the legislative enforcement might have had very little effect on the social norms among file sharers.

**Discussion**

Our preliminary results suggest that the majority of the posts responding to the question in the discussion topic title do not feel guilty for practicing illegal file sharing. Even though file sharing in Sweden during this time period went from a legal gray zone to being illegal, an increasing share of the posts state that file sharing is a right thing to do. Furthermore, our results indicate more buying of media, and possible changes in different rationales and practices for different genres and stakeholders. In all, the results show a multifaceted practice discussion which implies that IBCC is more much complex than just the question of illegal sharing of copyright protected files. These deeper aspects are challenging to investigate because it for obvious reasons might be hard to find active illegal file sharers and such explicit expressions of their perspectives like we did in this particular discussion topic.

The analysis presented in this paper is part of a larger study, in which illegal file sharing is investigated from different perspectives. Although not included in this study phase, when coding the data we noted a certain moral flexibility in the sense that different strategies and justifications seemed to be employed for different media genres. However, a deeper qualitative analysis is needed to further understand the complexity of current changes in file sharers’ justifications for what content to acquire through illegal file sharing, and what they choose to pay for – and why. Ultimately, this understanding would contribute to technical development of applications, systems and services for legal and fair IBCC by identifying different motivations, rationales, and justifications related to whether it is film, music, games, or programs that are consumed.
References

Is the Article Crucial to My Research? Evaluating Task-Oriented Impacts of Scientific Articles in Information Seeking

Jiepu Jiang, Daqing He, Shuguang Han, Wei Jeng
School of Information Sciences, University of Pittsburgh
jiepu.jiang@gmail.com, dah44@pitt.edu, shh69@pitt.edu, wej9@pitt.edu

Abstract

We propose a new aspect of evaluating scientific articles: crucialness, which refers to the state of articles being not only useful, but also scarce (difficult to be found or identified by scientists). Compared with the popularity-based metrics, crucialness may be a better metric in helping scientists' information seeking and use because it identifies scientists' difficulties in information seeking and reveals the crucial articles that may help scientists succeed in research. Some preliminary results are presented and discussed.

Keywords: crucialness; article evaluation; bibliometrics; information seeking; task.

Introduction

An abundance of metrics evaluate scientific resources based on their popularities, which can be measured by the number of citations (Garfield, 1972), the download statistics (Bollen, Van de Sompel, Hagberg, & Chute, 2009) and web 2.0 data (Jiang, He, & Ni, 2011; Priem & Hemminger, 2010). Studies (Wang & White, 1999) found these metrics can influence scholars’ information seeking and use. However, it remains unclear to what extent these metrics can help scientists and contribute to scientific discovery.

We argue that the popularity-based metrics may fail to solve two essential issues in scientists’ information seeking. First, the popular resources identified by these metrics (such as the highly cited articles and the journals with high impact factors) are usually well-known in the community. In practice, it is less likely that scientists need to find these resources, unless the scientists are new to the field. Instead, some useful but unpopular articles may deserve more attention and should be revealed to the community. Second, although an article’s popularity indicates high usage of the article, it is not necessarily that the article will be beneficial to the scientists who use it. In addition to popularity, we need to consider the actual effectiveness of using an article.

Assuming some important and useful articles are naturally difficult to be found, while some elite scientists have better abilities of finding these articles. Therefore, there should exist such a period of time that those articles are not noticed by the majority of scientists, while attract some elite scientists’ attention. We refer to such state of articles as crucialness, which implies not only high quality but also scarcity of the articles. Scarcity is a temporary state that the article is used limitedly and biasedly in the community. These articles are crucial to scientists’ research because they solve the majority of scientists’ difficulties in information seeking and shorten the gap between elite scientists and others. The crucial articles may finally turn into popular ones, as other scientists will gradually know these articles because of the elite scientists’ references to these articles and recommendations. Thus, revealing these articles can also speed up knowledge sharing among scientific community.

In this paper, we propose a calculable metric indicating the crucialness. Section 2 discusses the metric. Section 3 reports some preliminary results.

A Metric for Crucialness

A General Framework for Task-Oriented Metrics

We start with introducing a framework for evaluating the contribution of resources to the completion of a task. Without loss of generality, let \( X \) be the task of interest that makes users search and use resource \( R \). We assume the task will be either completed successfully or not (\( X \) and \( \bar{X} \)).


Copyright is held by the authors.
We define $\frac{P(R|X)}{P(R)}$ as a metric that indicates the crucialness of the resource $R$ for users’ successful completion of the task $X$: $P(R|X)$ is the usage rate of $R$ when users have successfully completed $X$; $P(R)$ is the overall usage rate of $R$ when users are trying to complete $X$ (no matter $X$ is successfully completed or not).

$$\frac{P(R|X)}{P(R)} \propto \frac{P(R|X) \cdot P(X)}{P(R)} = P(X|R)$$ \hspace{1cm} (1)

As in Eq(1), applying Bayes’ theorem, one can notice that the proposed metric is proportional to $P(X|R)$, the probability that $X$ can be completed successfully given we know $R$ has been used. Once we select the task $X$, $P(X)$ is a constant and can be ignored when comparing different articles. If we know using $R$ is a factor influencing $X$’s success, the metric indicates the contribution of $R$ for the successful completion of task $X$: $\frac{P(R|X)}{P(R)} > 1$ indicates that using $R$ is positively associated with the successful completion of $X$.

**Calculating Crucialness from Citation**

Based on the framework introduced in the above section, we can come to a metric of crucialness by defining $X$ and $R$ as follows:

\begin{itemize}
  \item $R$ An article used by scientists.
  \item $X$ Scientists write and publish new articles.
\end{itemize}

To make the metric calculable, we make the following settings: $X$ is counted as successfully completed if the published article is highly cited in the domain (referred to as HC articles); we say $R$ is used for completing $X$ if $R$ is cited by $X$. Practically, in a dataset, we select the top $k\%$ cited articles of a domain each year as the HC articles. We use $R$ and $X$ to characterize a task scenario that is ubiquitous among scientists: finding crucial articles that are helpful for conducting highly influential research (and consequently publishing highly influential and cited articles). Here we only consider the case that we have publication and citation data several years later than $R$ so that the metric is easily calculable. The in-time prediction of the metric is left as a future work.

The crucialness metric is calculated as Eq(2): $\frac{n_{HC}}{N_{HC}}$ and $\frac{n}{N}$ are the maximum likelihood estimations for $P(R|X)$ and $P(R)$.

$$\text{crucialness}(R) = \frac{P(R|X)}{P(R)} = \frac{n_{HC}}{N_{HC}} \cdot \frac{n}{N} \hspace{1cm} (2)$$

- $n_{HC}$ The number of times $R$ being cited by HC articles.
- $N_{HC}$ The total number of HC articles.
- $n$ The number of times $R$ being cited.
- $N$ The total number of articles.

Why is Eq(2) a metric indicating the crucialness of $R$? In a static dataset, when comparing the crucialness values of different articles, $N_{HC}$ and $N$ are constant and can be ignored in comparison. Thus, the metric in Eq(2) is proportional to $\frac{n_{HC}}{n}$, as in Eq(3). This makes certain connections between Eq(2) and our definition of crucialness. $n_{HC}$ and $n$ are citations of $R$ by HC articles and normal ones. If we consider authors of HC articles as an approximation of the elite scientists and authors of normal articles as the group of normal scientists, $n_{HC}$ and $n$ can roughly indicate the number of elite scientists and normal scientists found and used $R$ in their research.

$$\text{crucialness}(R) = \frac{P(R|X)}{P(R)} = \frac{n_{HC}}{N_{HC}} \cdot \frac{n}{N} \propto \frac{n_{HC}}{n} \approx \frac{\text{the number of elite scientists used } R}{\text{the number of scientists used } R} \hspace{1cm} (3)$$
The following two assumptions (Hₐ and Hₐ) make the crucialness metric, literally defined based on conditional probabilities, indicative of certain causal relationship between “using R” and “the successful completion of X”.

Hₐ Authors of HC articles (the elite scientists) and others differ in their abilities of assessing the values of scientific articles.

Hₐ Authors of HC articles (the elite scientists) and others differ in their abilities of finding useful scientific articles.

We assume that finding and assessing useful articles are difficult for the majority of scientists and require expertise, and the authors of HC articles (the elite scientists) do have better such expertise. Hₐ and Hₐ assume differences of scientists in their abilities of assessing and finding information, respectively. Let us also assume that scholars can freely search for articles and use articles; then, Hₐ and Hₐ result in the differences of article usage between authors of HC articles and the normal ones.

If we take Hₐ, n_HC and n indicate the judged quality of articles by scientists of different assessing abilities. In such case, articles with high crucialness values are those recognized high by elite scientists while low by normal scientists. If we take Hₐ, n_HC and n indicate the difficulties of finding the articles by scientists of different information seeking abilities. If the crucialness value > 1, a higher crucialness value suggests a larger gap between elite scientists and others in finding the articles. If either Hₐ or Hₐ is true, the metric indicates the crucialness of articles defined in our paper.

Some Properties of the Metric

The crucialness metric has many good properties that are beneficial for scholars’ seeking and assessing of academic information:

1. As discussed above, the metric is proportional to P(X|R). With proper assumptions, the metric measures the contribution of articles to scholars’ success in research.

2. As it is defined in the form of the ratio of two probabilities, i.e. P(R|X) and P(R), the metric is scaled by the productivity of the communities interested in R. Thus, articles from fields of different levels of citation rates can have comparable crucialness values (for example, articles from history and physics). This unique property makes the crucialness metric helpful for scholars when related articles come from multiple topics that have different levels of citation rates. In such cases, articles identified by popularities may be biased to those from topics with high citation rates.

Unsolved Issues

Although the crucialness metric has numerous useful properties, in this initial work, it is limited in its calculation methods:

1. We solely use citation frequency as the indicators for HC articles.

2. We solely use “citing” as evidences of whether scientists used one article or not.

3. Apparently, we need to wait several years until we can aggregate enough citations to calculate the crucialness value.

4. The two probabilities in Eq(2) are estimated by maximum likelihood estimation, which can be inaccurate if a small sample size is given (e.g. for articles with very low citation frequency). Such technical issues make it only possible, at current stage, to calculate the crucialness value for highly cited articles.

Solutions to these issues, especially effective in-time prediction of the crucialness values, may finally make the metric suitable for practical use. These issues are left as further works.

Preliminary Results

We select 39 journals from library and information science field from web of knowledge (WoK) “Information Science/Library Science” (IS/LS) category. In JCR 2009, the WoK IS/LS field includes 65 journals; we remove the following journals: (1) journals from information system domain; (2) journals from domain specific informatics (e.g. GIS, medical informatics). We remove these journals because the topics and articles related to these journals may largely use articles from other disciplines, which can result in inaccurate calculation of the crucialness values in our experiment. Articles for the 39 journals from 1954
to 2011 are collected. Only articles with “DT” field in WoK database equals to “article” or “proceeding article” are used in our study (other articles may include editorials and book reviews etc.).

In calculation of crucialness values, \( k \) is set to 20 (so that the top 20% cited articles each year will be selected as HC articles). We further normalize the citation frequency by topics. We calculate topic model of articles using LDA (we use both articles’ titles and abstracts). We tested for different numbers of topics to be extracted, and find 50 topics are appropriate for our study: fewer topics (e.g. 20) is usually not enough to identify fine-grained topics, and more topics (e.g. 100) will lead to insufficient sample size for our study. We further manually analyze and label the 50 topics, of which 24 are enough to topics to be extracted, and find 50 topics are appropriate for our model.

Table 1 shows citations and crucialness values for the top 10 cited articles from 1995 to 2010 in the dataset. As showed in the table, highly popular (cited) articles do differ a lot in their crucialness values, which may suggest their different values for scientists in information seeking. Some of the widely cited articles have even lower than 1 crucialness values, indicating these articles are popular but may not contribute to scientists’ success in research.

Our study suggests not all popular articles should be presented to scientists, and argues the usefulness of current popularity-based metrics in helping scientists’ information seeking and use. As we discussed, the proposed metric has many good properties related to scientists’ information seeking and use. However, the metric is not yet mature enough to be applied to practical systems due to the lack of time calculation methods, which will be the focus of our further studies.

<table>
<thead>
<tr>
<th>Cite</th>
<th>Crucialness</th>
<th>Year</th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>654</td>
<td>1.67</td>
<td>1995</td>
<td>S. Taylor; P. A. Todd</td>
<td>Understanding Information Technology Usage: A Test of Competing Models</td>
</tr>
<tr>
<td>628</td>
<td>0.38</td>
<td>2003</td>
<td>V. Venkatesh; M. G. Morris; G. B. Davis; F. D. Davis</td>
<td>User Acceptance of Information Technology: Toward a Unified View</td>
</tr>
<tr>
<td>441</td>
<td>1.21</td>
<td>1995</td>
<td>D. R. Compeau; C. A. Higgins</td>
<td>Computer Self-Efficacy: Development of a Measure and Initial Test</td>
</tr>
<tr>
<td>360</td>
<td>1.79</td>
<td>1995</td>
<td>D. L. Goodhue; R. L. Thompson</td>
<td>Task-Technology Fit and Individual Performance</td>
</tr>
<tr>
<td>351</td>
<td>2.26</td>
<td>2000</td>
<td>B. J. Jansen; A. Spink; T. Saracevic</td>
<td>Real Life, Real Users, and Real Needs: a Study and Analysis of User Queries on the Web</td>
</tr>
<tr>
<td>348</td>
<td>0.49</td>
<td>2003</td>
<td>W. H. DeLone; E. R. McLean</td>
<td>The DeLone and McLean Model of Information Systems Success: A Ten-Year Update</td>
</tr>
<tr>
<td>338</td>
<td>1.14</td>
<td>1999</td>
<td>D. W. Bates; J. M. Teich; J. Lee; D. Seger; G. J. Kuperman; N. Maluf; D. Boyle; L. Leape</td>
<td>The Impact of Computerized Physician Order Entry on Medication Error Prevention</td>
</tr>
<tr>
<td>325</td>
<td>1.38</td>
<td>1999</td>
<td>E. Karahanna; D. W. Straub; N. L. Chervany</td>
<td>Information Technology Adoption across Time: A Cross-sectional Comparison of Pre-adoption and Post-adoption Beliefs</td>
</tr>
<tr>
<td>310</td>
<td>1.38</td>
<td>1999</td>
<td>H. K. Klein; M. D. Myers</td>
<td>A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems</td>
</tr>
<tr>
<td>274</td>
<td>1.17</td>
<td>2000</td>
<td>R. Agarwal; E. Karahanna</td>
<td>Time Flies When You’re Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage</td>
</tr>
</tbody>
</table>
References


**Instructional Strategies for Autistic Adults Learning Video Game Design**

**Demetria Ennis-Cole**  
University of North Texas  
Demetria.Ennis-Cole@unt.edu

**Princess Cullum**  
University of North Texas  
Princessmcullum@hotmail.com

**Gary Moore**  
nonPareil Institute  
gary@ntipx.org

**Abstract**

A diagnosis of Autism Spectrum Disorders (ASD) indicates that an individual has a complex neurological disorder which impairs communication, behavior, and social interaction. ASD persists throughout an individual's lifetime; many adults with ASD are very knowledgeable and skilled, but they tend to be unemployed and underemployed because of difficulties interacting, communicating, and securing positions for which they are qualified (Chappel & Somers, 2010; Hendricks & Wehman, 2009; Nesbitt, 2000). This poster describes instructional strategies that match the information processing needs of adults with ASD. The poster is based on findings from a qualitative pilot study designed to identify best practices for teaching adults with ASD video game design. Data from semi-structured interviews with adult learners revealed direct, interactive, and experiential learning as instructional preferences along with the elimination of stressors (competition with others, limited time, and test taking).

**Keyword:** adults, ASD, instruction, technology, Autism

**Introduction**

Autism is a spectrum of disorders; several conditions are included in the spectrum: Autistic Disorder, Asperger Syndrome or Asperger's Disorder, and Pervasive Developmental Disorder- Not Otherwise Specified (PDD-NOS). According to the CDC, 1 in 88 children have a diagnosis of ASD (Autism Society of America, 2012). A great deal has been written about the needs of children with ASD. Volumes of scholarly literature exists on early intervention, teaching strategies, needs of families, efficacious therapies, technology tools, and special programs for children, but little has been written about the needs and issues of adults in the spectrum. Because ASD persists over the lifetime of an individual, children with ASD will become adults with ASD, and they will need some level of support to ensure a healthy, productive, and fulfilling life. Many individuals with ASD are able to use technology and are motivated by it. The use of technology in intervention planning is gaining momentum and producing promising results (Wainer & Ingersoll, 2010). Data from a qualitative study of adult learners with ASD reveals needs in instructional delivery and confirms underemployment. These findings should be considered by institutions, educators, and employers. The wisdom gathered from semi-structured interviews with adults with ASD can be used to inform the scholarship of teaching and learning.

**Goal of the Study**

The goal of this study was to explore practices at a training facility for adults with ASD to determine the instructional strategies used to teach video game design. Seven adult learners with ASD provided their perspectives on their experiences at the institute, their preferences in instructional environments, and their employment experiences. The research questions are as follows:

1. What instructional strategies are used to teach video game design to adult learners with ASD?
2. What is the employment status of interviewees with ASD?

---


Copyright is held by the author/owner(s).
Theoretical Framework

Social Constructivism is the theoretical framework for this study. In this theory, learners build unique and individual understanding based on their prior knowledge and experience. Learning takes place at the fringe of what is already known. Novice learners collaborate with more experienced learners who model outcomes, beliefs, procedures, and behaviors. Over time, the novice learns and applies the concepts modeled. In Social Constructivism, interaction is critical for learning content; interaction between beginning and more advanced learners supports practice and develops a climate for sharing knowledge.

Method

In an attempt to understand the instructional strategies and work experience of adults with ASD, contact was made with a technology institute for adults with ASD. The organization teaches design engine features and tools, map creation, and 3D Modeling. A request to conduct face-to-face interviews with students of the institute was made and granted. Subsequent University IRB approval was sought and granted. Students were selected by the institute and interviewed by the primary researcher. A diverse group of seven adult learners was selected based on their articulation ability and willingness to talk with the researcher. Two participants were African-American Males, one was a Caucasian Female, three were Caucasian Males, and one was an Asian-American Male. The institute verified an ASD diagnosis for each participant; all individuals at the institute are required to have valid testing from a licensed professional before admittance. Before the interviews began, the institute reviewed all questions.

The interviews lasted between 25 to 40 minutes and took place in a comfortable and quiet environment. Questions about employment, learning with technology at the institute, previous education, future goals, and class structure were asked. The final question was open-ended and asked if the interviewee had any additional information to share. All student interviews were recorded, transcribed, and printed.

The primary researcher conducted follow-up interviews with each student as a member check to ensure validity and gain additional information. Once interviews were verified, the primary and secondary researches utilized peer review to ensure reliability and validity of transcript data. The researchers worked together to identify and agree on themes, expand and group themes into overarching categories, and record the frequency of each theme.

Results

The interviews provided a wealth of information. Students' responses indicated that they prefer Direct, Interactive, and Experiential Instruction. Students at the institute are heavily involved in Direct Instruction through demonstrations and explicit teaching. They learn to manipulate 3D Models and game engines through modeling. A more advanced student or instructor walks them through features of the tools and explains how to use the tools effectively. Coaching, explanation, and supervision are used to deliver direct instruction that is individualized to meet the needs of each student. Comments from 57% of the students indicate that their learning is “self-paced,” based on their “needs and aptitude,” and “one-on-one.”

In addition, students’ comments indicate that they prefer interactive instruction. They mentioned “face-to-face instruction,” “follow-up by their instructor (both electronically and face-to-face),” “encouragement,” “feedback,” and “positive praise” as essential characteristics of a learning environment. Students interact directly with their instructor and they also interact with other students in a variety of ways. The most significant interaction occurs in Production Teams where students work together to learn to develop products and learn workplace skills. One student said, “Well it feels like I’m a part of the workforce, and I actually contribute something big.”

Experiential learning is used in several ways at the institute. Students report that they are able to “learn newer technologies,” and “expand and learn more” through their experiences and through their interaction with tools and people. They have the opportunity to create games and play them when they need downtime or have a break. Their interaction with other gamers creates additional experiences for individual learning. One student indicated that he enjoyed the “The fact that I’m surrounded by gamers
like me -- The whole organization revolves around Autism and video gaming. This microworld is easier than things in the real world."

Participants in this study were also asked about their employment experiences to determine whether or not these matched findings in the literature - adults with ASD tend to be unemployed or underemployed because of social, communicative, or behavioral challenges (Chappel & Somers, 2010; Hendricks & Wehman, 2009; Nesbitt, 2000; Taylor & Seltzer, 2011). Underemployment was reported by students prior to attending the institute. One adult was working at Walmart (14.2%), two (28.6%) were employed – one delivering papers, and one at a Taco Shop. The remaining students (57.2%) were living at home or attending community college. One student had worked as an Engineer, but was no longer in that profession. Since attending the institute, three of the interviewees (42.8%) have been hired by the institute as full or part-time instructors. The institute hires its advanced students to deliver instruction. These students are proficient using the software, they understand ASD, and they are able to deliver instruction effectively. The institute eliminates underemployment for some adults with ASD.

Conclusion

Many adults with ASD are very capable and technology savvy. Their educational experiences need to contain direct, interactive, and experiential instruction in order to help them reach their full potential. More programs and services for adults with ASD are needed to help them develop high-tech skills; more employment opportunities and training are needed also.

References

Fearless Cards: 
Computer Training for Extremely Marginalized Populations

Ricardo Gomez, Ivette Bayo, Philip Reed, Cherry Wang
University of Washington
Corresponding Author: Ricardo Gomez, rgomez@uw.edu

Abstract

We present a novel design for very basic computer literacy training for extremely marginalized populations in industrialized countries, such as immigrant day laborers in the US. In addition to lacking technical skills, day laborers experience strong emotional barriers such as fear, anxiety and lack of self-confidence, in addition to limited English language skills. Through an iterative process of design and testing, we developed the Fearless Cards, a set of basic, easy-to-use instructions to help extremely underserved populations overcome their emotional barriers and learn to use computers and the internet to help them improve their lives. Sample decks of Fearless Cards are available with the Poster Session for examination, testing and feedback.

Fearless Cards are important because they help reach the deep pockets of exclusion and allow people who are extremely marginalized to break the emotional barriers that prevent them from starting to use computers and the internet, they help to strengthen their sense of empowerment and self-worth, and help promote stronger self-confidence and self-efficacy.

Keywords: computer training, emotional barriers, marginalized populations, community development, immigrants

Copyright is held by the author/owner(s).
iDiscover: Inspiring Youth to Pursue STEM

Daniel Knopp  
U of Maryland, Baltimore County  
Dknopp1@umbc.edu

Olivia Green  
University of Pittsburgh  
Ocg3@pitt.edu

Clea Counts  
University of Miami-Ohio  
Countscm@muohio.edu

Abstract

This program was an experiment started in 2011 to look into the ways that we can diversify the STEM pipeline by targeting a group of middle-school students in Baltimore, Maryland. We also introduced Information Sciences principles during our presentation, but were limited by the near-complete lack of understanding about Information Sciences by the students at this public school. We worked with several sources to develop a set of handouts and materials that we hoped would assist the students in pursuing STEM and Information Sciences in college. Surveys were administered to gather preliminary data on our efforts. We then created a Wiki (idiscover.wikispot.org) where further development of the project from other parties can continue.

Keywords: STEM, information, Pittsburgh, diversity, education

Research Question

How can we increase minority participation in Science, Technology, Math, Engineering (STEM), and Information Sciences (IS)?

Introduction

Students are often more enthusiastic about STEM when they are exposed to the field through group-activities (Rhodes et al. 2011). We feel that we are unique to the diversification efforts in Information Science because we are unique in bringing the IS & STEM focus earlier in the pipeline to reach children at a more critical age. We realized we needed a program that addresses needs of minority students such as economic concerns and a need for real-world and group experiences. We hoped to continue the Information Science field’s efforts to gain inclusion.

We set out to design a program (we named iDiscover) that met the following goals:

• Provide information about scholarships.
• Provide information about tutoring services.
• Provide information about real-world activities related to STEM (Anakwe & Greenhaus, 2000).
• Encourage students to engage in STEM by providing literature on college requirements and career possibilities (Thiry et al. 2011).
• Build a curriculum around these findings, and created handouts for the students.
• Conduct a Pre/Post survey to gauge iDiscover’s effectiveness.

Method

We first gave a presentation orienting the students with the nature of Information Sciences and STEM. We then connected this to real life examples by showing, in the presentation, jobs available for IS and STEM graduates, as well as how these jobs will grow in availability. We then discussed what Maryland colleges want to see from prospective students to make them competitive for STEM programs. After that, we provided resources on tutoring and financial aid to help the students achieve their goals.


Copyright is held by the author/owner(s).
We passed out the documents in person as well as posting them on a Wiki. We also provided them with a step-by-step path to college from 6th-12th grade, including admissions goals as well as academic and financial ones. We predicted that we would have at least a 50% increase in the intention to pursue mathematics, since we focused on the career possibilities of encouraging students to engage in STEM by educating them on college requirements and career possibilities the most (Thiry et al. 2011).

We also predicted that the interest in STEM would rise by at least 50%. We predicted that the rise in interest and knowledge about information sciences would rise by only 25%, since this was not covered as thoroughly in the presentation.

**Process**

Daniel Knopp, Clea Counts, and Olivia Green created a presentation to pitch the idea of a STEM program to the Baltimore City Public Schools (BCPS), looking to help refine the program and set target groups. The BCPS agreed with our findings from our interviews and research that 8th grade would be a good grade to teach this program. They stated that mathematics must be emphasized the most, because that is the biggest topic in city schools right now, so we adjusted the curriculum accordingly. Eventually we were placed with a middle school in southwest Baltimore for a 50-minute segment (one class period). We had built a curriculum for after-school and longer, so we simply consolidated the information. Daniel Knopp taught a selected group of 28 students who had applied to STEM charter schools, but did not get in. This was to show that we had a group who at least had some interest in entering a STEM discipline, but may have been discouraged by not getting into the STEM high schools.

**Statistics from iDiscover**

**Qualitative Observations from iDiscover**

- Conducted at a middle school: Located in inner city Baltimore. 54% Female, 46% Male.
- Very motivated set of students and staff.
- Some technology was archaic but still present.
- Students were most interested in financial aid and scholarships, intrigued by the summer programs. Very excited when distributing the resources we gathered for them.
- Some students had difficulty grasping the concept of Information Science. Became more interested when I spoke to some students individually about the interdisciplinary nature of IS.
- Held a discussion with the students where they suggested how information on social sites could be organized.
- School administration stressed that the students needed tutoring resources in mathematics.

**Quantitative Results**

- Compared these two questions:
  - Before completing the iDiscover Program, did you have a good understanding about STEM?
  - After completing iDiscover, are you more aware of STEM and Information Sciences? (From post-survey)
  - **Highly significant (.000).**
  - Before completing the iDiscover program, rate how likely you were to pursue each of these majors in college?
    - Interest in Math had a **significant (.002)** increase. Interest in furthering reading comprehension skills also increased **significantly (.034).**
    - After completing iDiscover, rate how likely you are to pursue each of these majors in college?
      - One questionnaire pair that approached being significant, but was not quite (.057) was the question related to pursuing a major in Information Science. There was almost a significant increase, given that higher scores mean more interest.
Question 3 on the post survey is the same as the pretest Question 4. *(Are you interested in going to college after graduating high school?)* This could be viewed as a check on reliability of the question. It also revealed a significant improvement in knowledge based on our training, significant at (.001).

**Conclusion & Expansion**

The middle-school was impressed with iDiscover and asked us to return and present the program to more students. In future presentations, we should focus on STEM more than Information Sciences because of an expected lack of education amongst middle school students in Information Sciences.

Several members of i3 cohort were interested in using the gathered materials for their own versions of the program in the U.S. Virgin Islands and Pittsburgh public schools.

In the future, we should expand to cover Information Sciences more thoroughly. We could look at studies detailing STEM outreach efforts to further refine our efforts. We wanted to focus on each aspect of STEM individually, but were limited by the time allocated for our presentation. In the future we could also incorporate knowledgeable teachers and guests. There are a few organizations in the Baltimore area specializing in STEM outreach that are willing to provide guest speakers.

**References**


Dawn Best, (Principal in New York City Public Schools), interview by Daniel Knopp, 07 15, 2011.

Deb Philcox, (Director of Community Engagement for Baltimore City Public Schools), interview by Daniel Knopp, 07 08, 2011.


DataUp: Enabling Data Stewardship for Researchers

Carly Strasser
University of California Curation Center, California Digital Library

carly.strasser@ucop.edu

Abstract

The move towards digital data is ubiquitous across all domains in academic research (Interagency Working Group on Digital Data, 2009; Carlson, 2006; C. L. Borgman, 2009; Faniel & Zimmerman, 2011; C. L. Borgman, 2008), and these data can be made available and distributed more quickly than ever before. This is often called the data deluge, and is a phenomenon that has been explored in the traditional academic literature (Carlson, 2006; Faniel & Zimmerman, 2011; C. Borgman, Wallis, & Enyedy, 2007), as well as in several major media outlets (Editors, 2010; Pollack, 2011; G Bell, 2009).

Among the most pressing problems associated with the data deluge is good data management: how does one handle the huge volume of available information effectively and efficiently to solve important problems? Knowledge of good data management techniques and software development lags behind the progression of the data deluge. Consequently, although researchers of all fields are faced with huge volumes of data from increasingly diverse sources, they do not have the skills to handle their data sets. This challenge is amplified by the fact that research data are seldom shared, re-used, or preserved (Nelson, 2009; Tenopir et al., 2011; LeClere, 2010). There is a growing awareness among practitioners and funders that this situation represents inefficient use of research dollars, missed opportunities to exploit prior investment, and a general loss for the scholarly community (Editors, 2009). Michener, Brunt, Helly, Kirchner, and Stafford (1997) described the loss of valuable data and insight about those datasets as “information entropy”. This loss of information is becoming increasingly worrisome as data management practices improve very slowly, while the volume of data grows exponentially.

Recognizing that most Earth, environmental, and ecological scientists use spreadsheets at some point in the life cycles of their data, the California Digital Library partnered with Microsoft Research Connections and the Gordon and Betty Moore Foundation to create a tool for Microsoft Excel that would encourage and enable good data stewardship practices. To optimize the tool, we first identified the needs of the community via surveys of researchers. We found that, on average, researchers had poor data management practices, were not aware of data centers or metadata standards, and did not understand the benefits of data management or sharing. We used the survey results to compose a list of desirable components and solicited feedback from the community to prioritize potential features. The result of this effort was a document outlining the requirements for DataUp; these requirements were made publicly available for comment on the DataUp blog, and were provided to developers who then created the DataUp software.

The resulting DataUp tool (dataup.cdlib.org) facilitates documenting, managing, and archiving tabular scientific data. It comes in two forms, both open-source: an add-in for Excel and a web-based application. The add-in operates within the well-known program, Excel; the web application allows users to upload tabular data to the web-based tool in either Excel format or comma-separated value (CSV) format. Both the add-in and the web application provide users with the ability to (1) Perform a “best practices check” to ensure data are well formatted and organized; (2) Create standardized metadata, or a description of the data, using a wizard-style template; (3) Retrieve a unique identifier for their dataset from their data repository, and (4) Post their datasets and associated metadata to the repository.


Copyright is held by the author/owner(s).
Although there are hundreds of data repositories available to researchers for data archiving, the majority of scientists are not aware of their existence or how to access them. One of the major outcomes of the DataUp project is the ONEShare repository, created specifically for DataUp. Users can deposit their tabular data and metadata directly into the ONEShare repository from within the tool, allowing for seamless data archiving within the researcher’s current workflow. An added advantage of ONEShare is its connection to the DataONE network of repositories (Michener, 2009). DataONE links together existing data centers and enables its users to search for data across all participating repositories using a single search interface. Data deposited into ONEShare will be indexed and made discoverable by any DataONE user, facilitating collaboration and enabling data re-use.

CDL envisions the future of DataUp to be directed by the participating community at large. Code for both the add-in and web application is open source¹ and participation in its improvement is strongly encouraged. Interested developers can expand upon and increase the tool’s functionality to meet the needs of a broad array of researchers. To facilitate this effort, a list of requested improvements, bugs, and feature requests is maintained by CDL². Although the target audience for the tools that result from the DataUp project will be Earth, environmental, oceanographic, and ecological scientists, we believe that any tools developed will be adaptable to other research communities, such as the social sciences.

Keywords: data management, research data, Excel, services, software tools

References


LeClere, F. (2010, August). *Too many researchers are reluctant to share their data.*


1. http://bitbucket.org/dataup/main

The 'Forever Problem': Nuclear Waste as Information

Chris Heaney
University of Washington
Information School
chubbles@uw.edu

Abstract

Nuclear waste repositories harbor materials that are dangerous for tens of thousands of years, far longer than the normal human planning time span. Disposal facilities such as the Waste Isolation Pilot Plant and the halted Yucca Mountain plant offer an uncommon look into the history of long-term transmission and preservation of information. This poster explores concepts relating both to nuclear waste as information in and of itself, communicating through its emissions, as well as ways in which nuclear waste necessitates long-term, human-driven preservation of information. Furthermore, it probes issues regarding the design of systems for information transfer over long time periods, relations between various recent attempts to communicate across great temporal distances, and ethical issues surrounding obligations to warn future generations of hazardous materials.

Keywords: deep time, information ethics, long-term communication, nuclear waste, WIPP

Introduction

The United States government currently operates a deep underground storage repository for transuranic nuclear waste materials in southeastern New Mexico called the Waste Isolation Pilot Plant, or WIPP. This facility (active since 1999) and the planned Yucca Mountain facility in Nevada for spent nuclear fuel (halted in 2010) have been closely watched by nuclear regulatory agencies outside the United States, many of which have begun planning or constructing deep geological repositories for their own waste. WIPP and Yucca Mountain (if it becomes operational) are legally obligated, according to Environmental Protection Agency regulations, to establish methods to deter human intrusion into the repositories for a 10,000-year period following their closure (Chapman 2003).

A number of federally commissioned reports were authored in the 1980s and 1990s addressing the issue of how one goes about planning to warn far future societies about the dangers of radioactive materials buried deep underground (Human Interference Task Force 1984; Hora 1991; Trauth 1993). These papers were authored by a diverse crowd of scholars and intellectuals, drawing from physics, materials science, archaeology, anthropology, history, communications, linguistics, semiotics, and futures studies. Seemingly at once eminently practical (indeed, perhaps even morally necessary) and patently absurd, the reports range from matter-of-fact discussions of granite marker weathering to futuristic scenarios essentially indistinguishable from science fiction.

Significance

Comparatively little attention has been paid to nuclear waste disposal as a species of information problem, specifically the problem of information transfer and preservation over long time periods. There is a sense in which nuclear waste is itself information; it communicates time, through the regularity of decay, and danger, through bodily harm. In both cases, however, the waste needs interpretation; it is

doi:10.9776/13302
Copyright is held by the author/owner(s).
cryptic and insidious in its dispatches, necessitating human intervention to keep it from communicating solely via the loss of human life. To this end, myriad mechanisms for transmitting and preserving information regarding the waste's size, location, composition, date of placement, and hazards have been proposed. These range from multilingual stone obelisks, to worldwide archival depositories, to priesthood-like social institutions, to folk-tale-like cultural implantations. The timescale involved encourages the generation of uncommon methods of information preservation, drawn from contemplation and a long view of history rather than from cutting-edge engineering (Friedman 2010).

Technological advances of the last century have seen a movement toward ever faster and more easily accessible information devices and media, but concomitant with speed and accessibility are the threats of ephemerality, obsolescence, and atrophy. Information today goes more places and does it quicker, but it is in some ways more vulnerable to the ravages of time than ever before. The problem of information transfer over long time periods is not new, but its manifestations have changed across eras and civilizations, and these reflect the concerns, aspirations, and obsessions of their originating individuals, institutions, and cultures. The modern world presents a schizophrenic view, at once obsessed (in some corners) with information preservation and heedless (in other corners) of the necessity, or desirability, of attempting to communicate beyond the next few years.

In some ways, nuclear waste storage facilities resemble other types of modern information transfer and preservation projects, such as time capsules (Jarvis 2003), genetic and seed banks, Stewart Brand’s Clock of the Long Now (which seeks to build a mechanical clock which will tick for 10,000 years) (Brand 1999), fringe projects such as the Scientology storage vaults in California and New Mexico, and even the plaques and records attached to interstellar spacecraft (Sagan 1978). They can be framed as “memory institutions” in reverse; the markers, barriers, and knowledge storehouses that will warn the future about WIPP and its kind are at once memorials, monuments, and archives, but ones meant to spark repulsion rather than invitation, fear rather than curiosity (Foote 1990; Trauth 1993).

Furthermore, such projects offer an opportunity to probe ethical responsibilities toward future generations – after all, the placement of radioactive materials can be a literal matter of life and death. These repositories serve as a springboard to consider whether, if we can communicate with the far future, we should, and what we should attempt to communicate; and whether we should attempt to communicate over ‘deep time’ (Benford, 1999) even if we think we will not be successful. Being as the regulatory period specified by the United States government – 10,000 years – is roughly as long as the timespan of recorded history to date, and given that the material will actually be dangerous for far longer than this (plutonium-238, for instance, has a half-life of 24,000 years and so will be significantly dangerous for some 100,000 years) (Chapman 2003), nuclear waste is essentially a “forever problem”, permanently an information issue, and one with very real consequences.

Conclusion

The poster seeks to address the above concerns along with visual representations of markers, facilities, and information systems suggested, designed, and/or adopted for the waste storage units and related projects. It is intended to be a contribution to the history and philosophy of information, as well as an attempt to theorize, and promote thinking about, a particular species of information preservation that has attracted scant consideration from the information field specifically. Future work includes further historical research and analysis of documents relating to nuclear waste, such as government reports; hermeneutics-inspired (Gadamer 1989) attempts to derive fruitful theoretical frameworks for understanding ‘deep time’ projects; and critical analysis of ethical standpoints regarding intergenerational ethics as they relate to information transmission projects.

References


Abstract

Research questions were posed to examine users’ requests for information, beyond a keyframe, and certain influences thereof when interactively retrieving video. Laboratory-style search experiments employing 28 users from the field of science education were conducted. Users were asked to attempt six pre-designed search topics. All actions of the users and topic durations were recorded, and a post-search questionnaire was used to collect data about the topics and users’ experiences and impressions of the experimental system. Results showed that users regularly requested information about a clip, beyond a keyframe, while searching, not browsing, and that requests varied significantly across certain topic categories devised for this study. Users’ requests for information about a clip were negatively correlated with post-search satisfaction ratings, search times, and the number of actions performed. Findings also indicated that users' requests were not associated with a search topic's representation of the experimental domain. Such findings give researchers additional insight about user interface design for video digital libraries.

Keywords: video retrieval, user interaction, user interface design, keyframes, digital libraries

Introduction

User-centered video retrieval research is significant considering the complex makeup of digital video, which lends itself to a vast range of possibilities and considerations for designing retrieval tools like user interfaces. Video’s multidimensional makeup, granular structuring, effects of the "semantic gap," along with a collection’s scope, envisioned users, and/or intended domain(s) can all influence designs of user interfaces and promote variations in video representations across different video retrieval systems (Albertson, 2010; Christel, 2008a; Huurnink et al., 2010). Ding et al. (1999) concluded that hybrid video representations, or video surrogates, comprising both textual and visual information, were generally more effective than textual or visual representations alone. Christel (2008b) noted the benefits of visual surrogates specifically, with regards to browsing, in that they enable users to quickly scan the collection. Amir et al. (2003) also generalized variations in the usefulness of different video components, and stated that visual information is useful for browsing while transcript (textual) data is most applicable for search. Such previous findings are significant and certainly inform the design of video retrieval systems (i.e. digital libraries) and user interfaces. However, questions remain, especially those that examine users’ requests, use, and perceptions of such individual components of video in order to aid retrieval and selection.

Research Questions

Research questions were posed to better understand users’ requests for information beyond a keyframe, when presented, through the user interface, as the initial representation of a video clip. The general overarching question posed as part of this analysis included:

• What factors influence users to request information from a video retrieval system beyond a keyframe; or, conversely, what are some of the circumstances where users are satisfied and confident with a keyframe alone?
Other more-precise exploratory questions were developed to measure the influence of certain experimental factors on information requests by the users. The specific experimental factors examined here included some of those related to information needs and actual users. For example:

- How does the makeup of video search topics influence users' requests for additional information from a video digital library?
- Does a search topic's representativeness of a collection's domain influence users to seek more information?
- Do users who are more or less familiar with the video information being sought request more or less from the system?
- As search sessions prolong, what are the effects on users' requests for additional information about a clip and subsequently user satisfaction?

Findings from these questions, which examine user interaction and judgments, will further inform interface design for interactive video retrieval systems, such as video digital libraries.

**Methodology and System**

Science Education was the experimental domain (or context) used to explore these research questions. Teachers and aspiring teachers (college seniors) were recruited; a total 28 users participated in this study. Laboratory-style interactive search experiments were designed and carried out that asked each user to attempt different experimental search topics. Topics were categorized based on the number of subtopics and their incorporation of visual and/or textual needs, resulting in the following types:

- Basic: Contained one subtopic, based on either visual or textual information alone.
- Complex: Contained multiple subtopics, all of which based on textual or visual information alone.
- Combination: Contained one subtopic involving both textual and visual information.
- Combo-complex: Contained more than one subtopic involving both textual and visual information.

Two of each type of topic were created. Users completed a total of six (out of eight) experimental search topics. Individual topics were systematically ordered; each was given at each spot in the rotation an equal amount of times throughout the search experiments.

The interactive search experiments were conducted using a prototype video retrieval system designed to search and browse the NASA K–16 Science Education Programs. This collection comprises several NASA educational series, such as NASA Connect, NASA SciFiles, NASA Why?Files, and Destination Tomorrow, with production dates spanning 2000 to 2006. The experimental system, or, more specifically, its user interface, offered both visual and textual search features and various browse categories; however, for experimental purposes, a keyframe was the initial/default way of viewing results.

A questionnaire was given after each search, which asked users to rate their familiarity with the search topic, the representativeness of the experimental topics with "real" search topics from Science Education, topic difficulty, and satisfaction with the user interface and search results. The lead researcher on this study monitored the search experiments and manually recorded all actions performed by each user and search times. The number of times each interface feature was used was tallied, and Web logs helped validate observations. Different quantitative analyses, including mean analysis, Pearson's correlation, and ANOVA tests, were conducted.

**Findings**

Results showed that users' requests for additional information about a clip, beyond a keyframe, were significantly correlated (r=.407, p < 0.01, N=112) with keyword searches, in positive manner, but not video browsing; that is, as the number of keyword searches increased, so did users' requests, and vice versa. This finding was reasonable, as keyword searches would likely be utilized more for semantic needs; therefore, users would want to assess where and how their search query corresponded to a video clip. Video browsing, on the other hand, is steered by access point; therefore, selecting a title, for example, combined with a keyframe may have been sufficient to indicate suitability of the returned clips. These findings give researchers insight about the design of user interfaces for new collections based on how they envision users accessing items.
Users’ requests for further information were shown to vary according to the different categories of search topics. A one-way ANOVA of the topic groupings yielded significant differences, F (3, 164) = 2.883, p < 0.05. Significant differences from an LSD post hoc analysis, also at p < 0.05, occurred between basic topics and both complex and combo-complex types of topics at \( (M = -1.524, 95\%\ CI [-2.75 -0.30]) \) and \( (M = -1.571, 95\%\ CI [-2.80 -0.35]) \), respectively. To generalize this finding, users requested more from the system about a clip for topics asking for multiple sources of information, but not for topics requesting both visual and semantic (textual) information as part of one topic. Significance of incorporating additional metadata in video search results to support users with more detailed needs is verified, but not necessarily for supporting users, such as general audiences, with predominately shorter basic needs.

A topic’s resemblance of the experimental domain did not produce significant associations with users’ requests for additional information. This finding suggested that such requests were more global of an interaction. However, further influences of the domain can be indirectly assessed by evaluating results in conjunction with knowledgeable users’ familiarity with the experimental search topics. Higher familiarity with these topics led to fewer information requests about a clip; that is, the more science educators knew what they were looking for, the less they examined video clips, as indicated by a significant negative correlation at \( (r = -0.224, p < 0.01, N=168) \). This finding suggests possible variations in user interactions according to collection type or scope, in terms of what type of additional information is being requested to support information needs of knowledgeable users.

Users’ requests for additional information about a video clip were also found to be positively correlated with the overall number of user actions \( (r = 0.602, p < 0.01, N=168) \), search time \( (r = 0.484, p < 0.01, N=168) \), and perceived topic difficulty \( (r = 0.364, p < 0.01, N=168) \). These findings demonstrated that the more time users spent attempting search topics, the more information they requested from the system’s database. This outcome is also supportive of the findings uncovered for a previous (second) research question, in that as needs increase in the amount of required information, so do users’ requests. With regards to addressing user satisfaction, or how added interactivity of the user impacted their perceptions, findings suggested that as user satisfaction increased, users’ requests decreased, and vice versa. Negative correlations at \( p < 0.01 \) were discovered between users’ requests for more information and both user satisfaction with the user interface and search results \( (r = -0.301 \) and \( r = -0.369, p < 0.01, N=168) \). This finding was intriguing in that users were more satisfied when they had ultimately retrieved less information from the collection for verifying a given topic was complete. Additional time having to explore the actual contents of individual video clips reduced satisfaction among users.

References


Meta-Scraping: Two Technological Approaches to Support Meta-analyses

Kim Nimon  
University of North Texas  
kim.nimon@unt.edu

Cornelia Caragea  
University of North Texas  
Cornelia.caragea@unt.edu

Frederick L. Oswald  
Rice University  
fred.oswald@gmail.com

Abstract

Meta-analysis is a principled statistical approach for summarizing quantitative information reported across studies within a research domain of interest. Although the results of meta-analyses can be highly informative for taking a broad conceptual and empirical approach to an existing body of research literature, the process of collecting and coding the data for a meta-analysis is often a labor-intensive effort fraught with the potential for human error and idiosyncrasy, as researchers typically spend weeks poring over journal articles, technical reports, book chapters and other materials provided by researchers in order to retrieve key data elements that are then manually coded into some form of a spreadsheet for subsequent analyses (e.g., descriptive statistics, effect sizes, reliability estimates, demographics, study conditions). In this poster, we identify two technological solutions to support the process of collecting data for a meta-analysis.

Keywords: meta-analysis, information extraction, machine learning

Introduction

Meta-analysis is a principled statistical approach for summarizing quantitative information reported across research studies within a domain of interest (see Hedges & Olkin, 1985). Meta-analysis is important to applied research as it synthesizes data from multiple studies to estimate the magnitude, consistency, and predictors of reported statistics. In the case of social science research, validity and reliability coefficients may be meta-analyzed in order to explain variability across studies for different tests and measures (cf. Schmidt & Hunter, 1977; Vacha-Haase, 1988). In information science, meta-analysis has been used to consider the effects of users’ cognitive ability on information visualization systems (cf. Chen & Yu, 2000).

Problem

Although the results of meta-analyses can be highly informative for taking a broad conceptual and empirical approach to an existing body of research literature, the process of collecting and coding the data for a meta-analysis is often a labor-intensive effort fraught with the potential for human error and idiosyncrasy, as researchers typically spend weeks poring over journal articles, technical reports, book chapters and other materials provided by researchers in order to retrieve key data elements that are then manually coded into some form of a spreadsheet for subsequent analyses (e.g., descriptive statistics, effect sizes, reliability estimates, demographics, study conditions).

Consider the case of a meta-analysis of the reliability estimates generated from the Organizational Commitment Questionnaire (OCQ). In support of the popularity of the OCQ, a Google Scholar search on the two seminal publications related to the OCQ — Mowday, Steers, and Porter, 1979; and Porter, Steers, Mowday, and Boulian, 1974 – results in 4,175 and 3,444 citations, respectively. After duplicate citations have been eliminated, the relevant article for each citation must then be imported from an electronic database or scanned in from a print journal. Then the process of examining each article to retrieve reported reliability estimates and other study parameters begins. Given the vast number of

Copyright is held by the author/owner(s).
collaborators is helpful to divide the work, it presents critical issues regarding the nature and articles to review, such a study would likely be conducted by a team of researchers. Although a team of reliability of the coding process. For example, psychometric checks for inter-coder agreement are critical. However, imagine the very real problem of critical pieces of information that were either not gathered or gathered inconsistently across research members. Coders can share the same blind spots or opinions of the data, where they may mistakenly agree on the wrong pieces of information that are present or absent. In other words, statistical indices of agreement will not capture all the shared omissions or mistaken consistencies in comparing codes. It would help to determine ways that improve the process and efficiency of coding information or meta-analysis.

Proposed Solutions

In this poster, we identify two solutions to support the process of collecting data for a meta-analysis. To explain these solutions concretely, we consider the aforementioned meta-analysis of the reliability coefficients for the OCQ, where each research article in the corpus of articles to analyze may contain information such as sample size, reliability estimate, and response rate. This information needs to be extracted and stored in a spreadsheet for subsequent meta-analysis. Examples of such information contained in two different research articles are “A total of 345 questionnaires were distributed in 16 clinics to administrative and medical personnel, and 200 usable questionnaires were returned for a response rate of 58%,” (Cohen & Vigoda, 1999, p. 395), and “One-hundred and forty-four questionnaires were distributed and 118 completed questionnaires were returned to the authors, for a response rate of 82%. Eleven questionnaires were incomplete, so data from 107 responses were used in the analysis” (Luttman et al., 2003, p. 118). Thus, from the above text, we are interested in a tool that would extract the number of usable questionnaires and the response rate.

Machine Learning-based Solution

Here, we consider techniques for extracting structured information from semi-structured or unstructured text. Specifically, we propose to employ both supervised and semi-supervised rule-based learning, commonly used in information extraction tasks (Mitchell, 1998). Rule-based learning involves manually identifying a set of rules from a document corpus and applying it to new documents to extract the desired information (i.e., via pattern matching). These techniques can make use of context information surrounding a token of interest from a text document in order to extract that particular token and store it in the appropriate field in the table (e.g., in the first italicized text above, the tokens of interest are the number 200 and the percentage 58%, which could be extracted and stored in a comma delimited file for subsequent analyses).

Based on a collection of already manually coded/annotated articles, we will build field-specific dictionaries, which comprise words or phrases that are most likely to be found or to describe a particular field of interest. For example, the dictionary specific to the sample size field could comprise words such as “usable,” “questionnaires,” “completed,” “responses,” “usable questionnaires,” and “completed questionnaires.” Second, to avoid the extraction of undesired information such as “Eleven questionnaires” in the second italicized text above, we will build “positive” and “negative” word lists for each field-specific dictionary, and analyze a window of 5 to 10 words surrounding a word in the dictionary to determine the occurrence of words in this window from both positive and negative lists. We will then define sets of rules based on the dictionaries and the word lists. For example, IF (DIGIT questionnaires, returned) THEN extract, IF (DIGIT questionnaires, incomplete) THEN not_extract.

Journal-driven Solution

In addition to extracting text of existing journal articles for quantitative information that is usable for meta-analysis, we also are urging the editors of journals to provide data in a format more usable for meta-analysis so that in the future, the extraction process and its associated guesswork is less necessary.

---

1 Note that we first use PDFBox, available at [http://pdfbox.apache.org/](http://pdfbox.apache.org/), to convert PDF research articles to text.
and perhaps even unnecessary. More specifically, we are working with a small number of journal editors, who in turn are discussing with their respective publishers about the feasibility of providing tables in journal articles (e.g., correlation or covariance matrices) as supplemental material for download. We expect that each editor will have a unique set of external practical constraints and internally motivated desires, such that the nature and format of the available data across journals might vary a great deal. We would argue that these forms of diversity are to be welcomed and not feared, because providing data is certainly better than the current state of affairs where no tables are provided and humans are required for all of the coding. Diversity should be welcomed because it is a key component to healthy evolution: Different data-gathering methods and formats can be compared across journals; hybrids across these different approaches might be created; and journals who happen to be lagging in the practice of tabling and data sharing can learn from their predecessors to create new formats that (at least to them) reflect improvements.

Discussion

Although meta analysts have traditionally relied on the power of individuals to extract information from articles, technology and data sharing should be used to improve this process. We expect that combining human intelligence with machine intelligence and data sharing will improve the reliability and validity of data that is collected from meta-analyses and make this line of research available to more applied researchers. Although the current designs proposed do not remedy the issue of unpublished studies, future research should consider implementing the designs proposed and evaluate how well the solutions perform in comparison to the process of hand-coding articles. It will be interesting to determine what level of semantic ambiguity, easily understood by humans, can be accurately coded via technology.

References


Abstract

With the growing international population of scholars in Knowledge Organization (KO), international academic journals are a source of academic communication among researchers. Even with the inclusion of these journals in international databases, there are still limitations due to language, culture, and interpretation. The purpose of this study is to develop a comprehensive list of topics that comprise KO in these respective countries. This study is a content analysis of two KO journals from each country: Brazil, South Korea, Spain, and the United States. The journals are limited to the last five years of each journal, 2008 to 2012. Through the content analysis, this study contributes to increase communication among academics in KO.

Keywords: knowledge organization, international comparative study, academic communication, content analysis

Introduction

In Library and Information Science, specifically in Knowledge Organization, academic journals are the main vehicle for academic communication among researchers. In spite of the efforts of internationalization and the inclusion of journals from different countries in international databases (such as: Thomson Reuters, Scopus, EBSCO etc.), there are still some limitations. These limitations in academic communication may include aspects involving language, culture, and different interpretations of a domain.

Researchers that work in countries with primary languages other than English tend to publish their research in their own language and its national journals. When the research from these countries national journals are included or referenced in international databases, the different terms used in the articles’ titles, abstracts, or keywords can cause confusion among researchers developing studies about the same topic.

Pettigrew and McKechnie (2001) conducted a content analysis of six journals in English with the aim to understand how Library and Information Science’s authors have used theories in their research. Freitas, Gabriel Junior and Bufrem (2012) have identified academic collaboration and traditions of research among Brazilian and Spanish authors in the Knowledge Organization area. The authors applied a content analysis and bibliometric analysis to the Reference Database of Journal Articles on Information Science and perceived that Brazilian authors have used Spanish authors as theoretical background. The authors also identified Spanish authors publishing in Brazilian journals as well as co-authorship among Brazilians and Spanish authors in Brazilian journals. Thus, “...it is already possible to observe a tendency towards a theoretical proximity of Brazilian and Spanish authors in the area, marked by the citations and the thematic relationships which they represent” (Freitas, Gabriel Junior & Bufrem, 2012, p. 223).

So, considering the problem of low communication among academic communities from different countries using different languages and different backgrounds, this project aims to:
a) Gain insight into the research topics conducted in these four different countries;  
b) Present the scenery of Knowledge Organization in order to demonstrate how these countries have talked about this domain; and  
c) Verify the differences and similarities on the topics in order to understand and identify problems of communication among the researchers of these four countries.

This project is an international comparative study that contributes to the improvement of communications amongst academics in Knowledge Organization by synthesizing and identifying research topics in Knowledge Organization across four countries. It may also suggest potential possibilities and opportunities for areas of research collaboration among international researchers in Knowledge Organization.

**Methods**

This project will apply the method of content analysis, as a domain analytical approach (Hjørland & Albrechtsen, 1995; Hjørland, 2002; Hjørland & Hartel, 2003; Talja, 2005). The literature that will be chosen for analysis are selected by the following criteria: 1) 5 years Highest Impact Factor indicator for the subject category of Information Science and Library Science in Social Science Citation Index (from Thomson Reuters via Web of Knowledge), and 2) research team’s expertise (purposive sampling).

For the selection of journals for Spain and Brazil, the highest impact factor of Social Science Citation Index in Information Science and Library Science subject is applied. For the selection of the North American journals, both criteria are used. The subject category, Information Science and Library Science in Social Science Citation Index includes journals such as Library and Information Science, Computer Science, Medical Information, etc. Some journals having high impact factors such as Journal of Computer-Mediated Communication, Journal of the American Medical Informatics Association etc. are not included to be analyzed for this project based on their specificity. Therefore, the research team filtered the list of journals based on expertise. Since there are no Library and Information Science journals published in South Korean in the Social Sciences Citation Index, two South Korean journals have been purposively selected by the research team.

Some countries may have specialized journals mainly focusing on KO issues like the Knowledge Organization journal in United States, whereas many other countries do not have a journal specifically dealing with KO issues. In order to balance the scope of data, KO specialized journals are not included in data collection but these can be used as a counterpoint analysis for further projects offering elements to compare the topics extracted from general journals and KO journals. Table 1 shows the selected journals for the study. The scope of data is limited to the last five years of each journal, 2007 to 2011. Since the project does not focus on the quantitative aspects like the frequency of research topics in KO, the different size of data from each journal is not considered.

**Table 1**

**Selected LIS journals from each country**

<table>
<thead>
<tr>
<th>Countries</th>
<th>Selected Journals (2007-2011)</th>
</tr>
</thead>
</table>
| Brazil         | *Perspectivas em Ciência da Informação*  
|                | *Transinformação*                                                                           |
| South Korea    | *Korean Society for Information Management*  
|                | *Korean Society for Library and Information Science*                                         |
| Spain          | *Revista Española de Documentación Científica*  
|                | *Profesional de la Información*                                                             |
| United States  | *Journal of the American Society for Information Science and Technology*  
|                | *Library & Information Science Research*                                                    |

For the data analysis, each article is coded by four researchers in order to identify whether the articles is related to KO or not. Intercoder reliability is measured and shows above 99% agreement among researchers. Keywords from the KO related articles are listed and compared among four countries.
Preliminary Results

This poster partially analyzed data from 2010 to 2011. Table 2 shows the most frequently used keywords from each country. There are many topics such as classification, indexing, or information seeking behavior, which are commonly studied in all countries. However, some keywords tend to be studied more often in a certain country. For example, a keyword, *Epistemology of Library Science* (or Information Science), only appear in Brazilian journals.

Table 2
Most frequently used keywords from each country

<table>
<thead>
<tr>
<th>Brazil</th>
<th>Spain</th>
<th>South Korea</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information science</td>
<td>Spain</td>
<td>Public libraries</td>
<td>Science</td>
</tr>
<tr>
<td>Knowledge management</td>
<td>Online journalism</td>
<td>Academic libraries</td>
<td>Web</td>
</tr>
<tr>
<td>Citation analysis</td>
<td>Web 2.0</td>
<td>Citation analysis</td>
<td>Model(s)</td>
</tr>
<tr>
<td>Digital library</td>
<td>Bibliometric indicators</td>
<td>Information literacy</td>
<td>Impact</td>
</tr>
<tr>
<td>Scientific production</td>
<td>Scholarly journals</td>
<td>School libraries</td>
<td>Information</td>
</tr>
<tr>
<td>Information science</td>
<td>Internet</td>
<td>Information seeking behavior</td>
<td>Retrieval</td>
</tr>
<tr>
<td>Co-authorship</td>
<td>Evaluation</td>
<td>behavior</td>
<td>Seeking</td>
</tr>
<tr>
<td>Bibliometrics</td>
<td>Information architecture</td>
<td>Metadata</td>
<td>Indicate(ors)</td>
</tr>
<tr>
<td>Capes digital library of</td>
<td>Online newspapers</td>
<td>User studies</td>
<td>Information-retrieval</td>
</tr>
<tr>
<td>scientific journals</td>
<td></td>
<td>Folksonomies</td>
<td>Behavior</td>
</tr>
<tr>
<td>Environmental education</td>
<td></td>
<td>Ontology</td>
<td>Communication</td>
</tr>
</tbody>
</table>

This poster shows preliminary results of KO research topics that have been studied in four different countries. However, it is too early to make any conclusion at this point. The project requires further analysis to discover the different research topics in the four countries. The research team expects to show more results during the conference.

References


Two Paths to Motivation through Game Design Elements: Reward-Based Gamification and Meaningful Gamification

Scott Nicholson
Syracuse University
School of Information Studies
Because Play Matters game lab
scott@scottnicholson.com

Abstract

One definition of gamification is the "use of game design elements in non-game contexts" (Deterding et al, 2011, p.1). Gamification is used to change behavior, to educate, or to motivate. Many corporate gamification systems rely upon rewards and a very thin layer of a game experience to engage people through points, levels, leaderboards, achievements, and badges. This type of reward-based gamification has become commonplace, almost to the point of being expected, in new social media and information-based applications. Reward-based systems can be appropriate to engage people in short-term activities or to teach people valuable skills. When used for long-term change, however, there are some significant concerns about reward-based gamification.

In one of the core books on gamification, Gamification by Design, the authors state that "once you start giving someone a reward, you have to keep her in that reward loop forever" (Zichermann & Cunningham, 2011, p. 27). In the marketing messages for gamification systems and online presentations about gamification, this warning is not typically given. Organizations are starting reward-based gamification systems unaware that they will need to maintain these rewards to maintain engagement.

The motivational theory behind his concern is based on Organismic Integration Theory (OIT), which is part of Self-Determination Theory. OIT presents how people integrate external motivation into their sense of self. If someone perceives an reward as an attempt to control behavior, then that external motivation will be attached to the rewarded behavior in a negative way to the person's sense of self. If those rewards are then taken away, then the person's motivation to engage in the behavior will also disappear (Deci, Koestner, and Ryan, 2001).

On the other hand, if someone finds meaningful connections between his or her own interests or their background and the desired behavior, then the chances are much greater that the person will integrate that behavior into his or her sense of self in a positive way. The goal of meaningful gamification, as developed by Nicholson (2012a), is to help users find meaningful connections with the underlying non-game activities, and using rewards only when truly necessary. In order to do this, multiple types of game and play experiences are needed, as the theory of situational relevance tells us that different users find meaning in different things (Schamber, 1994). Another method of allowing users to find what is meaningful is to allow the users to create their own goals and rewards within a gamification structure. A key design constraint is that the needs of the users need to be placed before the needs of the organization in creating a system for meaningful gamification (Nicholson, 2012a).

From an information science perspective, meaningful gamification is about using game elements to convey relevant information about a non-game context. Games can be used as information containers to communicate. As the user engages with the gamification system, he or she discovers information about the non-game activity that has been placed there by the designer. This can be done explicitly through text, voice, or other traditional ways of conveying information, or can be done implicitly through embedding information in the play of the game that causes moments of discovery and reflection. This concept of persuading through the play of a game is known as procedural rhetoric (Bogost, 2007).
The goal of this poster is to contrast the reward-based gamification model with the model of meaningful gamification. The poster will explain the basic concepts of gamification, present the two models of gamification, discuss when each type of gamification is appropriate, provide gamification and meaningful gamification examples, and point attendees to additional information about gamification. The poster will present Nicholson’s BLAP model of gamification (2012a), which focuses on Badges, Levels & Leaderboards, Achievements, and Points, and compare it to Nicholson’s recently developed RECIPE for meaningful gamification, which focuses on Reflection, Exposition, Choice, Information, Play, and Engagement (developed from 2012b).

Keywords: gamification, games, motivation, relevance

References

Smart TV as a Social Platform

Dong-Hee Shin  
dshin@skku.edu

Taeyang Kim  
sunneverstop@hanmail.net

Department of Interaction Science, Sungkyunkwan University

Abstract

With the conceptual model of interactivity, this study empirically investigates the effects of perceived interactivity on the motivations and attitudes toward Smart TV in Korea. A model is created to validate the relationship of perceived interactivity to performance, attitude, and intention. Further, the model examines the mediating roles of perceived interactivity in the effect of performance on attitude toward Smart TV. Empirical evidence supports the mediating role of perceived interactivity. Implications of the findings are discussed in terms of building a theory of interactivity and providing practical insights into developing a user-centered Smart TV interface.

Keywords: Smart TV, social interaction, moderating effects, TAM, TRA

Introduction

Smart TV (STV), a television set with integrated interactive Internet capabilities, is being rapidly developed around the world. The rapid growth of high-speed broadband connections is enabling the creation of a smarter TV by delivering content directly to consumer devices in the living room. STVs have adapted the popularity of app stores for smartphones and brought them to living rooms. STVs enable viewers to not only navigate programs live, but also check online content such as news, weather forecasts, stock market information, maps and games. It will be capable of television-commerce, Internet browsing, twitting, online social networking, chat, and other things. STVs present new opportunities to consumer electronics manufacturers for product differentiation and value creation through user-driven product innovation. Media companies will be afforded a direct path to consumers, delivering more choices and new types of content in an on-demand, personalized service.

With all the progressive movement, it is expected that the era of STVs will come at a much faster rate than had been expected, following the success of the smartphone era. STVs will be at the heart of all electronics goods and communications devices used in homes. Others even predict that the time for STVs will come within one or two years. These positive predictions argue that STVs will become a so-called hub at home by connecting computers and telephones in homes and automating electricity, tap water, home security and other home entertainment systems. For example, people will check their doorbell security cameras through smart TVs while seeing doctors offering a ubiquitous health system. Along with STVs, smartphones will keep evolving, serving complementary functions in those smart homes. Because smartphones will be connected with STVs, users will be able to be check what is going on in their houses even when they are out.

However, with all of the hype over STVs, a question is whether marketers really know how consumers truly feel about STVs. As with many rollouts, a key problem is that STV is still primitive. STVs in the average household are probably years away. One factor slowing the adoption of STVs is the fact that content is scarce and service is limited. This is consonant with Cesar and Chorianopoulos’ study [1], which examined interactive TV and gave explanation on why the full potential of interactive TV has not yet been realized. They examined how viewers interact with TV content and argued that three factors (content editing, content sharing and content control) are important in the success of interactive TV. The current problem of STV is lack of content and thus customers have rare opportunity to edit, share, and control contents. Because of the low interactive features, customer apathy seems to be increased.
Consumers are not so exciting about another form of technology that will come and go within a few years. Consumers have seen so many technologies and so many rosy promises come and go over a decade. They still clearly remember the advent of similar technologies which ended up a total failure such as Web TV, TiVo, and television commerce (t-commerce). More recent technological innovations related to TVs include IPTV and 3DTV, which have yet to become widespread. Some consumers object to upgrading again to STV sets after recently upgrading to costly multimedia technology such as digital TV and 3DTV. Although pricing has not yet been revealed, incoming STVs are generally expected to be costly, partly due to various applications. While there is widespread enthusiasm for STVs, skeptics consider STVs to be a gimmick or, at best, an immature technology.

Despite rising concerns over user acceptance and marketability, STV viability issues have been addressed in only a few industry reports. While it is expected that STVs will likely succeed due to their interactive services, it is still unsure how the interactivity will play out in the diffusion of STVs. This study examines consumers’ perceptions of STVs by proposing a research model that incorporates perceived interactivity (PI), perceived hedonic performance, and perceived utilitarian performance as enhancing constructs. For the methodology, the structural equation modeling approach, supported by AMOS software version 18, was applied to assess the empirical strength of the relationships in the integrated model. The findings should be of interest to both academics and industry. From a theoretical perspective, this study suggests a model to identify antecedents of user intention to accept STVs. The new model represents an improvement over previous technology acceptance research (like the Theory of Planned Behavior and the Technology Acceptance Model), because it integrates cognitive and behavioral attitudes as the primary factors of influence. These attitudes are driven by underlying beliefs and assessing the beliefs will improve STV forecasting. From a practical standpoint, the findings may be used to guide industry in selecting more effective strategies to attract STV consumers. The STV industry is facing the challenge of how to design STV services that are useful, valuable, enjoyable and most importantly, user-centered. However, interface designs and related elements are rarely examined in-depth in the specific context of in-situ user evaluations. By better understanding how motivational factors will impact adoption and consumer behavior, manufacturers, content providers, and programmers can obtain the insights they need to plan their STV strategies. The results of this study represent a set of guidelines to help STV industry and developers better understand how users develop their perceptions of STVs and how users contribute to ongoing adoption and usage. Inception. So far, Samsung, the current leader in the 3DTV field, has sold 90 percent of all 3DTV sets in the market and is expected to sell 600,000 of them in the first six months of 2011.

Definition and Current Trends

The Evolution of Smart TV

STVs aim to focus on being easy, fun, and useful. The first products out of the gate will be equipped with a graphical user interface that will allow users quick and easy access to a variety of content and applications, many of the moves enabled just by a single click. STV manufacturers also focus on developing personalized applications in entertainment, games, lifestyles, education, news and information, which could be easily downloaded to television sets.

So far, STV has been rapidly developed by Korean technology giants like Samsung and LG Electronics. The two are vowing to take a leadership role in the STV market, which is the buzzword in consumer electronics. Both companies introduced their first model STVs equipped with network access for Web-connected televisions at the IFA trade fair in Berlin. STVs will provide an all-in-one device for data-handling television. However, the television makers lack in areas such as software capabilities and content delivery. As such, STV manufacturers are seeking partnerships with movie studios, entertainment companies and other content providers to reduce the gap, while also seeking to use their manufacturing prowess as an edge.

Although STV has made great strides technologically, it has still many obstacles to overcome, particularly usability. For example, user interfaces and form factors will face major changes — instead of being fixed in the center of the living room, they will be incorporated into mirrors, walls and ceilings to embody a layout-free concept. Since STV is in its infancy, it is difficult to predict just what the STV landscape will look like in the near future. One thing is clear though, STVs should focus on user-centered design to become a mainstream phenomenon. In this regard, this study focuses on user attitudes and
behaviors related to STVs with a focus on the role of PI and its relation to other factors of STV adoption and usage.

**User’s Experience With STV**

TAM was inspired by the TRA, which argued that both the attitude towards an action and the subjective norm have an impact on behavioral intention that, in turn, affects people’s performance and action. TAM assumed that perceived usefulness (PU) and perceived ease of use (PEou) were major influences on an individual’s attitude toward using technology [2]. This study proposes an adaptation of the TAM that consists of social presence, perceived enjoyment (PE), perceived quality, and flow, along with the traditional components of TAM. Figure 1 presents the proposed Smart TV acceptance and use model. The utility of considering the TAM stems from the fact that 3DTV is technology-driven, as well as user-oriented. This model, based on the TRA, is well-suited to reflect the nature of Smart TV, because the model embodies the evolutionary progression of technology to become more agile, interactive, and easy-to-use. All the key drivers are defined and explained, and their relationships to acceptance of Smart TV are examined. Given the wide applicability of the TAM in emerging technologies, it is expected that the general causalities found in TAM are also applicable to Smart TV. In particular, the relationship between attitude and intention in IPTV (Internet Protocol TV) has been confirmed [3]. Thus, it is hypothesized that:

- H1: Attitude toward STVs is positively related to the intention to use STVs.
- H2: PI positively influences intention to use STVs.
- H3: PI positively influences attitude toward STVs.
- H4: PI positively influences perceived hedonic performance of STVs.
- H5: PI positively influences perceived utilitarian performance of STVs.
- H6: The higher utilitarian performance a user perceives of an STV, the more positive an attitude the user has about STVs.
- H7: The higher hedonic performance a user perceives of an STV, the more positive an attitude the user has about STVs.

**Study Design**

The survey method consisted of four phases. First, individual in-depth interviews were conducted with potential customers. Ten respondents were asked to explain their attitude and the experience of advanced TV such as 3DTV, HDTV and IPTV. Second, with the help of the STV manufacturers in Korea (Samsung), five focus groups of current and prospective users were organized, and group interviews were conducted in which groups of four to six individuals discussed how they currently use TV services and what factors would influence their use of STV services in the future. The goal of the individual interviews and focus group sessions was to test and validate the research model, to identify items missing from the model, and to gain a preliminary understanding of the factors that have an impact on usage behaviors.

Third, based on the focus group sessions, a final survey questionnaire was developed through several comment rounds of an expert panel consisting of professors, researchers, and STV experts. Prior to its use, the questionnaire was tested by administering a pilot survey among possible users who, in turn, provided a comprehensive review of individual responses to the pretest survey. Twenty undergraduate students participated, with tests given at three-week intervals. Prior to answering the questionnaire, they were strictly instructed to ask the experimenter any questions about questionnaire items that they did not understand. With these precautions, the possibility of participants filling out some questions without exactly understanding the content of those questions was eliminated. The wording of items was reviewed and modified by three marketing professors knowledgeable in quantitative research based on the pilot test outcomes.

The finalized survey was administered online. From July to October 2010, a web-based survey questionnaire was posted in the communities of several professional associations, and on blogs and forums devoted to smartphones, advanced TV services, HDTV, digital TV, interactive TV, and 3DTV. The survey included preliminary questions to ensure the respondents had a certain level of understanding of STV. Such questions, for example, include “Do you know about STV?” and “Have you used STV before?” A cover letter was attached to explain its purpose and to ensure confidentiality. By the time the survey ended, 1208 visitors had viewed it, and 342 questionnaires were submitted. Of the submitted
questionnaires, 13 were excluded because of incomplete answers, leaving 329 usable responses. Table 3 presents the sample demographics. The final sample reflects the general population interested in STVs. For the analysis of statistics, AMOS, a maximum likelihood-based SEM software, was used.

**Measurement Development**

The variables in the model are well established in the HCI, Information System and Communications literature. Prior to further study, a pilot test for measures was conducted. The participants indicated their agreement with a set of statements using a 7-point Likert-type scale (ranging from “strongly disagree” to “strongly agree”) drawn from previously validated instruments. The final scales used in this study consisted of 24 items, all of which included three items for each variable.

**Pretests**

A pretest was undertaken to examine the test-retest reliability and to construct reliability indices before conducting fieldwork. Thirty current and prospective users, who have interests and/or had had experiences with 3DTV and other similar services (e.g. HDTV, augmented reality TV, interactive TV, IPTV) participated in the two pretests at an interval of three weeks. After eliminating the measure items that failed in either the retest or the alpha test, Cronbach’s alpha was applied to identify poor item-to-total-correlation items. The alpha values ranged between 0.84 and 0.91, suggesting acceptable construct reliability. When theoretical models do not exist, these pretests are useful in the early stages of empirical analysis in cases for which the basic purpose is exploration.

**Data Analysis**

**Measurement instrument.** The reliability and validity of the measurement instrument were evaluated with AMOS 18 using the reliability and convergent validity criteria. The reliability of the survey instrument was established by calculating Cronbach’s alpha to measure internal consistency. All values were above the recommended level of 0.7 [4]. The convergent and discriminant validity of the model were examined using the procedure suggested by Fornell and Larcker (1981), who recommend measuring the reliability of each measure and each construct, as well as the average variance extracted (AVE) for each construct. The reliability of each item was examined according to a principle components factor analysis. It shows the results of this analysis, with varimax rotation on the original 24 items (four items were eliminated due to low loading). According to Hair et al. (1995), measurement of items loads highly if the loading coefficient is above 0.6. This analysis showed that most items had factor loadings higher than 0.7, which Fornell and Larcker (1981) considered to be very significant. Each item loaded significantly on its underlying construct (p<0.01 in all cases). Therefore, all constructs in the model had adequate reliability and convergent validity.

To examine the discriminant validity, this study compared the shared variance between constructs with the average variance extracted from the individual constructs. The shared variance between constructs was lower than the average variance extracted from the individual constructs, confirming discriminant validity. In short, the measurement model demonstrated adequate reliability, convergent and discriminant validity. Content validity is a characteristic of items that are representative and drawn from established literature [4]. The variables in this study that were derived from the existing literature exhibited strong content validity. The wording used in the questionnaire appears similar, so that the responses can be highly inter-correlated. In addition, a correlation analysis of Pearson’s R (correlation coefficient) shows an acceptable level of correlation among variables.

**Structural model.** A test of the structural model was performed using the AMOS procedure, and a maximum likelihood-based SEM software was used. Table 2 shows the estimates from the structural modeling. The overall fit of the model is satisfactory, with all of the relevant goodness of fit indices greater than 0.90. The GFI is 0.95, the AGFI 0.91, and the TLI 0.91. Similarly, there is no evidence of misfit, with the RMSEA showing a very satisfactory level of 0.067, which favorably compares to the benchmarks by [5], who suggest that values of 0.06 or more reflect a close fit. The standardized RMR was also very good, at 0.027, well below the threshold for a good overall fit. Another positive test statistic was the normed chi-square value (a chi-square divided by degrees of freedom) of 1.98, a value that is appropriately below the benchmark of three, to indicate good overall model performance. Given a
satisfactory measurement of the model’s fit to the data, the path coefficients of the structural model were assessed.

Results

Structural Paths and Hypothesis Tests

A test of the structural model was performed using AMOS software. Table 4 shows the estimates from structural modeling. The overall fit of the model is satisfactory, with all of the relevant goodness of fit indices greater than 0.90. Chi-square statistics show non-significance in the models, indicating that the two models fit the data adequately. The GFI is 0.95, the AGFI 0.91, and the TLI 0.91. Similarly, there is no evidence of misfit, with the RMSEA showing a very satisfactory level of 0.067, which favorably compares to the benchmarks by [6] who suggest that values of 0.06 or more reflect a close fit. The standardized RMR was also very good, at 0.027, well below the threshold for a good overall fit. Another positive test statistic was the normed chi-square value of 1.98, a value that is appropriately well below the benchmark of 3, indicating good overall model performance. Given a satisfactory measurement of the model’s fit to the data, the path coefficients of the structural model were assessed.

To test the structural relationships, the hypothesized causal paths were estimated, and all seven hypotheses were supported. The results are reported and depicted in Table 5 and in Figure 3, respectively. The results support the proposed model well, confirming the key roles played by PI. All of the paths in the model are statistically significant. The results highlight the significant roles of PI in determining user attitudes toward STVs ($\beta=0.41$, $t=3.420$, $p < 0.01$), supporting H3. PI also has a significant direct effect both on PUP and PHP, which influence attitude significantly (H6 & H7). Whereas PUP and PHP had strong effects on attitude ($\beta=0.49$, $t=2.021$, $p<0.01; b =0.43$, $t=2.001$), the effect of attitude on intention was moderate or weak at the most in this model ($H1$, $\beta=0.24$, $t=3.120$); probably because the users want to confirm their intention with other factors, probably PI. Consistent with this inference, PI showed the highest impact, supporting $H2$ ($\beta=0.65$, $t=4.981$, $p < 0.001$). Approximately 58% of the variance in the intention of STVs was explained by the variables in the model ($R^2=0.581$). The $R^2$ of all endogenous constructs in the model exceeded 20%.

Overall, the model shows a pattern that highlights the importance of utility and hedonicity along with their antecedent, PI. However, the model also underplays the role of attitude, as compared with previous studies employing attitude. This implies that while the STV consumers might have a good attitude influenced by PUP and PHP, this does not automatically lead to intention. While consumers might cognitively perceive the excellent features of STVs well, they may not really intend to adopt or use it. They may want to personally ensure that the STV experience is a positive one and that programming is available. It may be inferred that there is a gap between attitude and intention in STVs. Psychological factors like interactivity can play a facilitating role between attitude and intention. This role has important implications in terms of theory and practice. Thus, further tests are necessary to uncover possible underlying effects.

Table 1
Summary of the hypothesis tests

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Path coefficient ($\beta$)</th>
<th>t-value</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1: Attitude $\rightarrow$ Intention</td>
<td>0.23*</td>
<td>3.120</td>
<td>Yes</td>
</tr>
<tr>
<td>H2: PI $\rightarrow$ Intention</td>
<td>0.65***</td>
<td>4.981</td>
<td>Yes</td>
</tr>
<tr>
<td>H3: PI $\rightarrow$ Attitude</td>
<td>0.41**</td>
<td>3.420</td>
<td>Yes</td>
</tr>
<tr>
<td>H4: PI $\rightarrow$ PHP</td>
<td>0.40**</td>
<td>2.001</td>
<td>Yes</td>
</tr>
<tr>
<td>H5: PI $\rightarrow$ PUP</td>
<td>0.34*</td>
<td>2.459</td>
<td>Yes</td>
</tr>
<tr>
<td>H6: PUP $\rightarrow$ Attitude</td>
<td>0.49**</td>
<td>2.021</td>
<td>Yes</td>
</tr>
<tr>
<td>H7: PHP $\rightarrow$ Attitude</td>
<td>0.43**</td>
<td>2.001</td>
<td>Yes</td>
</tr>
</tbody>
</table>

* $p<0.05$; ** $p<0.01$
Discussion

Findings from the Research Model

The goal of the study was to empirically test the STV interaction model in order to explain the development of individual behavioral intentions to interact with STVs. The results add to our understanding of user attitudes and intentions in a new HCI paradigm to clarify the implications for the development of effective STV services. Overall, the findings represent an extension to previous work on design characteristics and interactivity by showing interactivity leading to utility and enjoyment as an antecedent and a mediator of positive attitude and intention. Thus, interactivity has a hedonic component, supporting work by Cyr et al. [7] and Van der Heijden [8].

Among the constructs, the effect of PI shows a much stronger impact on intention than previous studies have indicated. While this may be partly because this study emphasized the effect of PI, the unusually high impact of PI suggests that STV users are more influenced by the interactive features in their decision to accept STVs than conventional TV or other advanced TV services (e.g., augmented reality TV, HDTV, IPTV). This finding implies that STV is not only perceived as a TV device for entertainment, but as a multi-tasking social, educational, informational and commercial tool. STVs will be likely to evolve as a social platform that unifies all functions and features and enables users to connect with online communities through STVs. That is, it redefines how people engage and interact across any application on STVs.

This inference is well suited to the performance value in the model. Along with the highly significant result of PI, the effects of PHP and PUP also show a much stronger impact on attitude than previous studies have shown [9]. It may be inferred that there was some kind of effects between PI and PHP/PUP. Together with interactivity and performance, it might very well be that interactivity with the TV increases user perception of utility and hedonicity.

Just as previous studies have consistently shown the importance of usability in technology adoption, this study confirms the importance of usability and further clarifies that usability can be greatly influenced by utility and hedonicity. These findings pinpoint a need for STVs to provide viewers with quality content, as well as interactive services. Although the issue of quality has emerged as a major factor in STV development, to date the research on this issue is quite sparse, especially from the perspective of user perception on interactivity. The perception of technology quality by user is a major factor for achieving market breakthrough. While many studies indicate the important role of quality in user adoption, not many indicate to what the specific nature of quality refers. In other words, quality can vary depending on different technologies. The specific nature of quality should be clarified according to technology. This study finds that the quality in the STV context refers to interactivity and further clarifies the components of interactivity. Most importantly, this study shows how such interactivity is related to other factors with different roles.

It has been argued that the most significant potential of STVs is high quality and versatility. As people turn increasingly to STVs for various services they formerly got from other sources, their expectations for those services will change. Those changing expectations will undoubtedly have an impact on the development of future STVs as multimedia tools for games, commerce and entertainment. In this study, the user perception of performance shows a much stronger impact on intention than previous studies have indicated [10].

Interesting findings can be derived from the insignificant relation between attitude and intention. This weak link is consonant with the insignificant role by attitude. All of the paths regarding attitude--the path of PUP to attitude (H6), the path of PHP to attitude (H7), and the path of PI to attitude (H3)--show weaker effects, as compared to other paths particularly regarding performance. As such, further tests are necessary to uncover possible underlying effects.

This study modified and extended the proposed model to test possible hidden effects. Previous studies have shown the moderating effects of variables in various IT contexts [10][11]. While recent technology acceptance research has revealed the importance of interactivity, research has failed to further investigate the matter of meaningful interaction beyond the single effect of controllability, responsiveness or personalization [12][13]. Thus, given the importance of interactivity in STVs, it is worthwhile to examine new roles played by PI because user perception of interactivity may greatly vary depending on different technologies. This study tested the mediating role of PI in the effect of PUP/PHP on attitude.
An Extended Model with Moderating Effects

To test mediating effects, this study used Baron and Kenny (1986)'s procedure, which has discussed four steps in investigating mediation:

Process (1): Show the correlations of the initial variables;
Process (2): Show the correlations of the initial variables with the mediator;
Process (3): Show how the mediator influences the outcome variable; and
Process (4): Establish that the mediator mediates the relationship of the initial variables.

With this procedure in mind, regression analyses were conducted specifically to examine (1) whether independent variables (PUP & PHP) significantly accounted for variance in the hypothesized mediator (PI), (2) whether variance in the mediator (PI) accounted for variance in attitude toward STVs, and (3) whether the relationship between independent variables (PUP & PHP) and the dependent variable (attitude) would no longer be significant once the variance in the dependent variable accounted for by the mediator was partialled out.

First, a model was fitted in which PUP/PHP was regressed on PI. The effect of the independent variable (PUP/PHP) significantly explained the variance in the hypothesized mediator PI (t = 4.42, F = 15.42, p < 0.001, r² = 0.10). This result suggested that the mediator PI was related to the independent variables PUP/PHP whose effects are supposedly mediated. Another regression model was run with attitude toward STVs as the dependent variable and the mediator PI as the independent variable. PI significantly accounted for variance in the dependent variable attitude (t = 5.56, F = 29.75, p < 0.001, r² = 0.24). A third regression model was fitted with attitude as the dependent variable and PUP/PHP as independent variables. A significant result was obtained (t = 3.23, F = 9.76, p < 0.001, r² = 0.15). Finally, a fourth model was conducted with attitude and PUP/PHP as independent variables. The effects of PUP/PHP were not significant (t = 1.00, p = 0.39) after the significant effect of the hypothesized mediator PI (t = 4.98, p < 0.001) was partialled out. Thus, PI is proven to be a full mediator between PUP/PHP and attitude

Implications For Theory And Practice

The contribution of this study is both theoretical and practical. With regard to theoretical advancement, the empirical findings demonstrate that employing interactivity and performance would be a worthwhile extension of TRA/TPB or TAM in STVs, as they were found to be influential in predicting the attitude and behavioral intention to adopt STVs. As an antecedent variable to performance, the role of interactivity is of importance in the STV context, because one of the limitations of technology acceptance literature is that it does not help us explain acceptance in ways that guide development, besides suggesting that system characteristics have an impact on perceptions of enjoyment and usefulness. Therefore, as many researchers argue [14], it is essential to understand the antecedents and the underlying effects of the key factors in order to explain eventual user acceptance and continuous use. With regard to TPB, behavioral intention can be viewed as an individual's underlying attitude, which ultimately determines behavioral intentions via attitude [15]. TPB has some limitations including a significant risk of confounding between attitudes and norms, since attitudes can often be reframed as norms and vice versa. Another limitation is the assumption that when someone forms an intention to act, they will be free to act without limitations. In practice, constraints such as limited ability, time, social or organizational limits, and unconscious habits will limit the freedom to act. The model in this study attempts to resolve these limitations. This study contributes to the literature on the TRA/TPB research by confirming that perceived usability can influence behavioral intentions through attitude. This study focuses on the user-centered perspective: how users perceive and use STVs and how STV factors play a role in the development of user attitudes. This can be a modest but heuristic contribution to research on STV acceptance, because previous research has studied them separately, leaving the relationship unclear.

Related to this implication, an intriguing and heuristic contribution of this study is the recognition of a directional relationship between PI, performance, and attitude. While interactivity is an embedded concept in multimedia technologies, the concept has apparently been under-researched. Aside from the relationship between PI and intention, it seems that the relationship is not apparent in the STV context. Given the unique nature of STV interaction, the relationship with PI and other factors should be clarified. Based on the findings of PI, this study reconceptualizes and redefines PI in the STV context. Previously,
PI has been measured with the capability to interact with system/technologies. In the STV context, the current notion of interactivity should be expanded to include new features provided by STVs. For example, one of the motivations of smartphone users is to be connected by constantly communicating with other users. With the advancement of ubiquitous technologies, the social expectation is that one is nearly always connected and reachable almost instantly via smartphones. It is considered that smartphones are the instrument of that connectedness. Assuming smartphone features apply to STVs, STVs will have similar functions, features, and user interfaces. Given these, the PI of STVs should be understood not only via responses and feedback, but also the feeling of connectedness enabled by such continuous interaction.

Applying this new notion of interactivity into STVs may render a real-time interaction system unnecessary. Rather, it is more effective to increase the user’s sense of connectedness or belonging. From this understanding, specific design features and characteristics of STV services can be utilized to achieve meaningful interactivity. For example, STV viewers might want to post comments on social networking sites when they are watching TV programs. This kind of continuous interaction loop may increase users’ sense of connectedness. Future studies should further investigate the complex interrelationships among interactivity, connectedness, presence, performance, and usability to clarify these intricate relationships. In particular, given its key roles, PI should be further extensively investigated in reference to the presence and flow in the STV context. For example, two people in geographically disparate regions could watch the same Internet-based show at the same time, and chat with each other about the program. For the two people, interactivity may mean a social presence. Interactivity can occur at many different levels and degrees of engagement, and it is important to differentiate between these levels.

Practical implications for the STV industry can be drawn in terms of strategies and new models for STVs. As STVs converge with other multimedia technologies, the concepts of interaction will be highlighted anew. The industry should focus on enhancing social presence through PI and increasing interactivity, user participation and involvement. The findings suggest that vendors should ensure that their device works and plays in accordance with user expectations and emotions. In addition, the findings imply that interactive content/services will be vitally important to the sales of STVs.

Although this study only employed STVs as the target technology, the research model captures the general characteristics of ongoing smart IT and hence the findings of this study can enhance our understanding of the factors leading to future intention of smart IT in general, which provides both utilitarian and hedonic functions with services such as communication, information, entertainment, and commerce. The focus on the experience of interactivity allows us to take a step closer to the design features of future smart IT. The results suggest the high relevance and great importance for smart IT to be designed with the capability of inducing an experience of interactivity in users. The more users feel interactivity in using the technology, the more they will perceive it to be of high utilitarian and hedonic performance and expectation exceeding, and the more they will feel satisfied with the technology and intent to continue its usage.

In conclusion, considering the ever-changing nature of smart technologies, this study elucidates motivations associated with STV acceptance and the implications for developing effective STV services. As users accept STVs as a new tool to communicate, collaborate, and entertain, industries should provide usable tools and platforms for users. STVs will be likely become an exciting and popular application in the near future. For STVs to become popular, developers need to understand individual perceptions and experiences concerning truly smart services.

References


Cardiac Rehabilitation Referral: An Information Gap

Claudia J. Gollop  Annie T. Chen  Wan-Ching Wu

gollop@email.unc.edu  atchen@email.unc.edu  wanchinw@live.unc.edu

University of North Carolina at Chapel Hill

Abstract

Cardiac rehabilitation (CR), consisting of exercise and diet modifications, has been proven to promote a healthy lifestyle that can extend life, particularly for survivors of cardiovascular events. Nonetheless, there is a long-standing concern regarding the underutilization of CR in general and especially by women. The American Association of Cardiovascular and Pulmonary Rehabilitation recommends all eligible persons be referred and participate in a CR program. However, participation and adherence to CR remain low. There appears to be CR referral information gap in many instances, and thus focus groups that were conducted investigated four main research questions. First, what did CR mean for these former CR participants? Second, how did participants find out about CR? Third, what kind of referral information was received? Last, what information should prospective CR program participants receive? The poster will present the background and motivation for the study, preliminary results, and discussion of these results.

Keywords: cardiac rehabilitation, referral, information needs

Introduction

It is a well-known fact that heart disease is still the primary cause of death in many parts of the world. According to the American Heart Association (2012), heart disease is the No. 1 cause of death in the United States today. Cardiac rehabilitation (CR), consisting of exercise, diet, and lifestyle modifications, has been proven to promote, maintain and even extend life, particularly for survivors of cardiovascular events. Grace, et al. (2002) states, “it is now well established that cardiovascular mortality can be reduced by approximately 25% when patients participate in a multifactorial CR program” (p. 127). Nonetheless, there is a long-standing concern regarding the underutilization of CR, especially by women (Daniels, et al. 2012; Sanderson, 2010).

The American Association of Cardiovascular and Pulmonary Rehabilitation recommends all eligible persons be referred and participate in a CR program. However, participation and adherence to CR remain low. Although a number of factors contribute to these low rates, one important aspect, which was investigated in this study, is whether a patient receives CR referral information upon discharge form the hospital at all. According to analysis of data from the American Heart Association’s Get With The Guidelines program, only 56% of eligible coronary artery disease (CAD) patients were referred to cardiac rehabilitation (Brown et al., 2009). There appears to be a CR referral information gap in many instances, and thus the authors set out to employ focus groups to examine cardiac rehabilitation referral experiences, particularly as they may impact the heart health of women. This poster will present preliminary results of the first part of this study, which involved focus groups with former CR participants.

This study investigated four main research questions. First, what did CR mean for these former CR participants? Second, how did participants find out about CR? Third, what kind of referral information was received? Last, what information should prospective CR program participants receive?

Acknowledgement: Eleanor M. and Frederick G. Kilgour Research Grant, School of Information and Library Science, University of North Carolina at Chapel Hill.


Copyright is held by the author/owner(s).
Methods

In the first part of this study, two focus groups were conducted at a fitness center that offered cardiac rehabilitation programs. All of the participants had previously participated or were currently participating in CR. In both sessions, five men and two women participated; thus, altogether there were 10 men and 4 women.

Results

With regard to the first question, participants came up with a number of themes concerning the meaning of CR. In addition to improving one’s physical condition, one participant said, there was an emotional component, as well as learning not to overdo it, and “knowing how hard to push for yourself later on when you leave.” The program helped them to change their lifestyle and improve their habits, not only in terms of exercise but also in other ways, such as diet. There was also a social component of CR programs that people found supportive and facilitating.

An important point was that some participants’ perceptions of CR and their need for it changed during the program. Some thought that it was “just all exercise,” but as they went through the program they realized that there was more – that “it [was] not just a temporary program, but it's something that’s long term and the class [would] get you started.” These experiences corroborate the results of a study by Janssen, De Gucht, van Exel and Maes (2012), in which they found that patients’ illness perceptions changed significantly over the course of cardiac rehabilitation. Among others, they experienced an increased sense of illness coherence, a greater sense of treatment control, and a lessened emotional impact of the disease.

As mentioned earlier, the literature shows that only a fraction of those who are eligible for CR are actually referred. Thus, a second research question was: how did those who participated in the focus groups find out about CR? Among the focus group participants, some were referred and some were not. In a number of cases, participants had responded to follow-up phone calls. Some individuals pointed out that personal referrals were helpful in getting people to enroll.

With regard to the third research question, the kinds of referral information received, subjects noted a variety, among them, brochures, booklets, personal visit in hospital, recruitment call (post hospital discharge) and no referral. The final question asked information prospective cardiac rehabilitation program participants should receive. The most salient point was that the information should explain what CR was all about. Many mentioned that it was not explained to them until after they had arrived for the first session. Many said that the information should attempt to address assumptions that people make about cardiac rehabilitation. Making material available through different modes of delivery such as print, video, smart phones and face-to-face contact would also be helpful. In addition to health care providers, the participants mentioned that former CR participants could talk to those potentially eligible to participate: “have graduates from rehab program, volunteer time at UNC Hospital or wherever and go and make personal contact with people who are facing going to cardiac rehab and tell personal experience.” Lastly, participants concurred that it took time to process information; thus, having it presented multiple times was helpful.

Discussion

The preliminary findings from this study contribute to existing work on CR in various ways. First, the results facilitate a deeper understanding of what prospective participants may need to know about CR that they may not know initially, what preconceptions they may have, and what helped them to re-conceptualize the meaning of CR. Next, the study investigated the channels through which individuals came to know of CR, and what modes of information presentation, the types of information, and delivery would be effective in promoting enrollment. This knowledge can be useful in the future for the development of interventions for promoting CR enrollment and adherence.

The results from this research are part of a larger study involving CR participants. Future focus groups will attempt to investigate perceptions of CR, suitable recruitment strategies and appropriate design of CR informational materials tailored to different demographic audiences. As stated earlier, women maintain lower rates of participation in CR programs than men. The demographic in this current study (four women and 10 men) seems to reflect the extant literature (Daniels, et al 2012). Nonetheless,
an emphasis on gaining a better understanding of CR referral practices relative to women with cardiovascular disease remains imperative.

The poster will present the background and motivation for the study, the preliminary results summarized above, and discussion of these results. The authors would like to close with a quote from one of the participants, who said that it was not just about exercise, "It was a whole total program. And that’s what I think didn’t come out here. It has to be explained. There’s more to it than just getting on the treadmill or running around the track."

References


Lights, Cameras, Action: Using Video to Create Instructor Presence in an Online Information Studies Course

Sharon Stoerger
Information Technology and Informatics
Rutgers, The State University of New Jersey
sharon.stoerger@rutgers.edu

Abstract

By definition, social presence involves making individuals feel connected and part of a community. Yet, creating social presence among the students and the instructor does not happen automatically and can be challenging to develop in online courses. In these learning environments, interactions are typically conducted via text-based and asynchronous media. The visual and auditory cues that are often associated with face-to-face interactions are absent in this form of communication. However, easy-to-use and inexpensive technologies that can foster presence are emerging. This project investigated the use of rich media to promote visual thinking and increase the instructor’s social presence. More specifically, instructor-created videos were used to enhance the sense of presence in a fully online course. Based on surveys, reflections, and unstructured follow-up interviews, the students responded favorably to the richer mode of communication and indicated that they felt a greater connection to the instructor, as well as their classmates.

Keywords: community of inquiry, engagement, epresence, media richness, online education, presence

Need/Rational

Instructor ePresence and Videos

Technology use and wireless access is on the rise, which can enable ubiquitous educational opportunities. This may create greater learning independence, but the technology itself is not a replacement for the teacher. Garrison, Anderson, and Archer (2000) developed a “community of inquiry” framework to emphasize the importance of teaching presence and its role in designing effective learning environments. In an online course, the primary mode of communication is lean, asynchronous, and text-based. There is typically no opportunity for face-to-face interaction among the students and the instructor. While this design is convenient for the students and offers opportunities for reflection, it can pose challenges in developing effective teacher presence.

Digital camcorders are small, versatile, and inexpensive. These mobile devices often enable the user to transfer the content with minimal effort to social media sites such as YouTube or Facebook. These affordances enable instructors to use these devices to create videos for educational purposes. The unpolished appearance of this video content gives it an authentic quality that conveys the presence of the teacher. Through these videos, students can be introduced to course content, while at the same time, see and hear their instructor. Further, the consumption and creation of video content is aligned with the concept of digital literacy, which is an important 21st century skill.

This poster will investigate an instructor’s use of rich media, in particular video content, to create a sense of presence in a fully online Information Studies course.

Copyright is held by the author/owner(s).
Description

This Project

For the purpose of this project, video content was created for a fully online Information Studies course. The majority of the students in this course were at the graduate level. Four types of videos were created by the instructor: 1) a personal introduction; 2) modular introductions; 3) video check-ins; and 4) modular “cool tools” introductions. The personal introduction was designed to set the tone for the course and to help students connect with the instructor. In this course, there were seven modules that were introduced via videos, as well. The purpose of these videos was to introduce the readings to the students, highlight key concepts, and present questions for the students to think about as the class moved into the discussions. Another type of video created for this course was a video check-in. While the focus of these videos was less on the content of the module, they provided the instructor the opportunity to clarify misunderstandings, address questions that surfaced during the week, and present related materials that could supplement students’ understanding. The function of the video check-ins was to enable the instructor to maintain a connection with students as they moved through the multi-week modules.

One of the assignments associated with each module involved a “cool tools” presentation. For each of the module topics, students were instructed to select a cool tool to virtually present to their peers. These presentations, which investigated the positives, negatives, and pedagogical practices of the selected tools, took place in the course management system (i.e., D2L) discussion forum. Videos created to introduce this portion of the course activities were shorter and more “playful” than the other types created for this course and emphasized creativity over content. At the end of the semester, the students were surveyed to assess the effectiveness and to determine ways to improve this type of content.

In the initial stages of this project, the intent was to use a Flip digital camcorder to record the videos for this course. This device was used to create videos used during the first weeks of the course. Throughout the semester, rich media alternatives (e.g., Jing and Animoto) were investigated as the instructor became more comfortable with the video creation process.

Evidence of Effectiveness

To determine whether the inclusion of this type of content had an impact on students, open-ended surveys (midterm and final) and a reflection activity (conducted during the last module of the course) were employed. Students were strongly encouraged, but not required, to complete the surveys. In contrast, the reflection was a required activity in the course. A selected group of students were also invited to participate in unstructured follow-up interviews to understand in greater detail their reflection comments and/or survey responses.

Data collected via the surveys and reiterated in the reflection activity indicated that more than half (n=13) of students were skeptical at first about the educational potential of videos posted to YouTube. As this student noted on the survey, she did not consider watching YouTube to be part of the learning process and described her dislike in this way: “I have a terrible habit of getting stuck in YouTube cycles that waste hours.” In addition, many of the students worked full-time as K-12 teachers or school librarians at institutions that blocked social media tools, which contributed to the mindset that online information is “bad.”

By the end of the semester, students recognized the value associated with the videos. One student described the educational potential of this medium in this manner: “I feel like this approach [the use of instructor-created videos] can be very effective, because the teacher’s obligation is to surround the student with resources.” Another student liked the ability to access the video content multiple times: “The videos are easily reviewable, and they ensure that students understand the concepts that they are being taught.” Students found it helpful to see the ways in which the technologies we discussed in class could be put into practice: “These videos were very inspiring for me, as it showed that our recently developed capabilities are already finding their way into the learning environment.”
Challenges

Technological changes are occurring rapidly. On April 12, 2011, Cisco announced that it planned to discontinue the production of the Flip camcorder (Grobart & Rusli, 2011). Because of the ephemeral nature of devices such as the Flip, alternative video production options were investigated. The screencasting site, Jing (http://www.techsmith.com/jing/), and the video slideshow maker, Animoto (http://animoto.com/), are examples of these sites. With Jing, full-motion videos that captured information presented on a computer screen, sound, and voice could be quickly created. Animoto transformed pictures, video clips, and music into videos. While the monetary costs of creating video content for students is low, there are other costs to consider. Many of these tools are easy to use, yet they still require a time investment on the part of the instructor to learn how to use them effectively. Another aspect of creating video that can take time is editing this type of content. Further, determining how to make the videos accessible to students is an additional point to consider. Instructors may want to create a YouTube Channel to post class-related videos. These videos could be uploaded to the course management systems (CMS), as well.

Applicability for Other Instructors/Institutions

At present, many mobile devices such as smartphones and tablets enable individuals to create short videos that can be uploaded to a wide variety of social media sites. The costs associated with creating an epresence via video include the recording device, the learning curve that accompanies the use of the video creation tool, the editing software, and the location where the videos will be housed. Nevertheless, many of the creation and publication tools are relatively easy-to-use. As a result, instructors who are not professional videographers can experiment with this type of rich media as a way to enhance their online presence and create a more effective learning environment.

References

http://www.fineminddesign.com/site/media/Readings/Week%2010/Lee(2004)_Presence%20Expli
cated_CT.pdf
Proposed Facets of a Serendipitous Digital Environment

Lori McCay-Peet
Dalhousie University
Halifax, Nova Scotia, Canada
mccay@dal.ca

Elaine G. Toms
University of Sheffield
Sheffield, United Kingdom
e.toms@sheffield.ac.uk

Abstract

Serendipity is broadly defined as an unexpected and valuable experience prompted by an individual’s interactions with ideas, information, objects, and phenomena; in digital worlds the challenge to date has been developing environments capable of facilitating serendipity. In this poster we present the early stages of the development of a means to measure whether digital environments such as social media sites and digital libraries have the potential to facilitate serendipity. Drawing from prior research and the initial results of an expert review, we outline the preliminary facets of a serendipitous digital environment (SDE) that will be used to develop a scalar instrument. This research will provide both a source for design elements to guide the development of digital environments that support serendipity and a tool for developers and researchers to assess the serendipitous nature of digital environments.

Keywords: human-computer interaction, serendipity, scale development, evaluation

Introduction

Serendipity provides enjoyment and wonder as well as interesting new directions and revelations for individuals that reverberate in their work and personal lives (McCay-Peet & Toms, forthcoming). Serendipity is often defined using terms such as unexpected, accident, chance, and luck and is used to denote: 1) a faculty or aptitude; or 2) a phenomenon or event (Merton & Barber, 2004). We define serendipity as an unexpected experience prompted by an individual’s valuable interaction with ideas, information, objects or phenomena. Based on this definition, information-rich, hyperlinked environments appear ideal for serendipitous experience. While some support this positive perspective (Thurman & Schifferes, 2012), others have expressed concern that the personalization of web content narrows the diversity of information and ideas we encounter (Zuckerman, 2011). These differing viewpoints underscore that digital environments are not neutral. A search engine designed for accuracy, for example, may not provide an environment as conducive to serendipity as a system that supports browsing, or makes semi-relevant results visible to the user.

But how can we ascertain how well specific digital environments support serendipity? While researchers could simply ask users, for example, “Does this social media site facilitate serendipity?”, a single question is insufficient to capture a multi-dimensional phenomenon (DeVellis, 2003). The approach we take is the development of a scale that assesses whether the environment has the critical elements to facilitate serendipity. Applying this well-documented approach (see DeVellis; MacKenzie, Podsakoff, & Podsakoff, 2011) the following initial steps were taken

Acknowledgements: Special thanks to the international set of researchers who participated in the expert review to date: Lennart Björneborn, Jannica Heinström, Stephann Makri, and Borchuluun Yadamsuren. Initial research was supported by grants to Toms from SSHRC, CFI, and the Canada Research Chairs Program while she was at Dalhousie University, Halifax, Nova Scotia, Canada and the work continues to be supported by a NCE GRAND project. McCay-Peet was awarded a SSHRC Doctoral Scholarship to support the research.

Copyright is held by the author/owner(s).
1. explore features of the environment that research suggests are associated with serendipity;
2. examine features through a digital lens, identifying facets of a serendipitous digital environment (SDE);
3. develop multiple items for a scalar questionnaire that captures the conceptual space of a SDE; and
4. assess the content validity of the facets and items.

Prior Research

Prior research suggests that some people may be more likely to experience information encountering than others (Erdelez, 1999) and personality traits such as extraversion may influence how often individuals experience incidental information acquisition (Heinström, 2006). Likewise, some environments, such as information grounds, are conducive to opportunistic discovery of information (Pálsdóttir, 2011). Björneborn identified ten serendipity dimensions, or features of the physical library space that prompt divergent behaviour; for example, unhampered access, diversity, and explorability. Based on Björneborn’s dimensions, a preliminary scale was developed to explore these dimensions in a digital environment (McCay-Peet & Toms, 2011). Participants (N=123) browsed an experimental information system and responded to a 20-item scalar questionnaire. Five factors that characterized a SDE were identified through exploratory factor analysis: enabled connections, introduced the unexpected, presented variety, triggered divergence, and induced curiosity. As these findings were based solely on Björneborn’s research in one physical setting, that work needed to be replicated for digital environments, and validated. We undertook to do that assessment and the following section explores the characteristics of a potentially serendipitous digital environment. This is the first stage of the scale development process – the conceptualization of the construct (MacKenzie, et al, 2011).

Conceptualization of the Construct

What makes one environment more likely to facilitate serendipity than another? In semi-structured interviews with 12 professionals and academics, we extracted 15 examples of work-related serendipity (McCay-Peet & Toms, forthcoming). We found that the conceptual space of a serendipitous environment is marked by the presentation of its content as well as the content itself. These findings, together with findings from prior research suggest five potential facets of a SDE

1. **Enables exploration:** A user’s assessment of the degree to which a digital environment supports the unimpeded examination of its information, ideas or resources (e.g., McCay-Peet & Toms, forthcoming; Björneborn, 2008).
2. **Trigger-rich:** A user’s assessment of the degree to which a digital environment is filled with a variety of information, ideas, or resources interesting and useful to the user (e.g., McCay-Peet & Toms, forthcoming; Björneborn, 2008; Sun, Sharples, & Makri, 2011).
3. **Enables connections:** A user’s assessment of the degree to which a digital environment exposes them to combinations of information, ideas, or resources that make relationships between topics apparent. (e.g., McCay-Peet & Toms, forthcoming; Björneborn, 2008; Sun, Sharples, & Makri, 2011).
4. **Highlights triggers:** A user’s assessment of the degree to which a digital environment actively points to or alerts users to interesting and useful information, ideas, or resources using visual, auditory, or tactile cues. (e.g., McCay-Peet & Toms, forthcoming; Björneborn, 2008; Campos, & Figueiredo, 2002; Rubin, Burkell, & Quan-Haase, 2011).
5. **Leads to the unexpected:** A user’s assessment of the degree to which a digital environment provides fertile ground for unanticipated or surprising interactions with information, ideas, or resources. (e.g., McCay-Peet & Toms, forthcoming; Björneborn, 2008; Rubin, Burkell, & Quan-Haase, 2011).

With the initial conceptualization of the construct of a SDE complete, this study’s focus is currently on the development of a measure of a SDE.
Development of a Measure of a SDE

Based on the proposed five facets of a SDE, items that can be used to assess the “serendipitousness” of a digital environment were developed using a deductive approach (Hinkin, 1998); the definition of each of the facets guided the generation of their respective items. Rules for scale item development were followed, e.g., keep language clear and simple (DeVellis, 2003). For example, the following item was generated from the Enables exploration facet definition: “[The digital environment] is easy to explore”. When administering this questionnaire, researchers will be able to replace “[the digital environment]” with the specific digital environment being tested. Participants would rate their level of agreement to these items on a 5-point agreement scale. For each facet, 7-10 items were developed, for a total of 43 statements.

An international set of researchers who have conducted research in the area of serendipity and related constructs were invited to participate in this review of the facets and items. The expert review (still in progress) is designed to improve the content validity of the SDE scale. Those who responded were emailed a questionnaire containing the facets and items. They were asked to respond to the following:

1) Are the facets sufficient, definitions clear, and do they adequately capture a SDE?
2) Are the items clear and capture the gist of the facet definitions? Can you suggest other items?

While careful attention is being paid to all responses, particular weight is being given when two or more experts make similar comments. The decision to accept or reject all suggestions will ultimately be that of the scale developers (DeVellis, 2003). To date, four experts have responded, with four more expected. Preliminary results suggest that the five facets are sufficient, though minor changes are being made to improve clarity. There are currently 35 items in the revised questionnaire, 7 items per facet, but more changes to the facets and items are anticipated based on incoming reviews.

Conclusions and Future Work

We have presented the results of the preliminary stages of scale development in the creation of a tool to assess how well digital environments serendipity. This initial stage suggests that there are five facets of a SDE: 1) Enables exploration, 2) Trigger-rich, 3) Enables connections, 4) Highlights triggers and 5) Leads to the unexpected. Input from expert reviewers will strengthen the content validity of the scalar questionnaire. The next steps in the scale development process will include further assessment of validity and scale evaluation and refinement (MacKenzie, et al., 2011). Once completed, this tool will be of use to researchers and practitioners who endeavor to build or augment current digital environments or to assess the efficacy of existing environments for their “serendipitousness.”

References

Large-Scale Digital Library User Searching:
What Role Does Domain Play?

Oksana L. Zavalina  
Oksana.Zavalina@unt.edu  
Department of Library and Information Sciences  
College of Information  
University of North Texas

Elena V. Vassilieva  
Elena.Vassilieva@unt.edu

Abstract

This poster presents findings of an exploratory comparative analysis of user search queries in three large-scale digital libraries: two domain-specific functioning in the domains of STEM education and US history, and one domain-independent. This study measured search query lengths and frequencies, and categorized search queries into ten search categories based on content analysis. Results suggest that domain-based differences (i.e., differences in user searching between digital libraries representing different domains) are more substantial than differences in user searching between domain-specific and domain-independent digital libraries. Domain-based differences in distribution of search categories between search query length and search query frequencies are statistically significant. These findings may have implications for design and evaluation of large-scale aggregations of digitized materials.

Keywords: search queries, comparative analysis, domain analysis, transaction log analysis, content analysis

Introduction and Problem Statement

Digitization of information objects of cultural, historical, and educational value has been actively supported by US federal funding agencies such as Institute for Museum and Library Services (IMLS) and National Science Foundation (NSF). In addition, these agencies have supported large-scale digital libraries that aggregate hundreds of digital collections consisting of millions of digitized items, and provide centralized access for this wealth of information. Three of the largest digital libraries of this type in the USA have been:

- IMLS-funded Digital Collections and Content (IMLSDCC) with 650 digital collections
- IMLS-funded Opening History (OH) with over 1500 digital collections focusing on US history.
- NSF-funded National Science Digital Library (NSDL) with 120 science, technology, engineering, and mathematics (STEM) education digital collections.

IMLSDCC is a domain-independent digital library aimed at a general audience and covering a wide range of subject areas and disciplines; OH and NSDL are domain-specific, i.e., are created for distinct audiences. User base for domain-independent information systems tends to include more novice

Acknowledgements: This research was supported by the University of North Texas Office of Research and Economic Development and Texas Center for Digital Knowledge. Authors thank developers of the Opening History and IMLSDCC for providing access to search logs used in the study.

Copyright is held by the author/owner(s).
than expert users while the audience of domain-specific information systems normally includes a higher proportion of domain experts. It was observed in numerous studies (e.g., Allen, 1991; Bates, 1972; Connaway, Johnson, & Searing, 1997; Hembrooke et al., 2005; Hsieh-Yee, 1993; Marchionini et al., 1993; Wildemuth, 2004; Zhang, Anghelescu, & Yuan, 2005) that information seeking behavior, including search query length, frequency, specificity etc., depends to a large extent on searcher’s level of knowledge both on a specific search topic and in the broader subject domain.

To ensure seamless intellectual access to rich pools of digital content, that is accumulated in these large-scale digital libraries, by their intended user communities (e.g., broad general audience, historians or STEM educators), the design of their respective information systems, including description of information objects in metadata records, choice of search options etc., should be informed not only by general user tasks of finding, identifying, selecting, and obtaining information (IFLA, 2008) but also by the information needs and search approaches specific to users’ respective domains, and levels of users’ domain knowledge. Previous studies of web searching discovered, for example, that humanities scholars most often include in their search queries personal and geographic names, chronological and discipline terms (Bates, 1996); water quality researchers frequently use topical, geographical, and format or genre search terms, and occasionally—chemical formulas, dates, names, and URLs (Nowick & Mering, 2003); medical researchers’ prevailing search query types include laboratory/test results, disease/syndrome, body part/organ/organ component, pharmacological substance, or diagnostic procedure (Natarajan et al., 2010).

Transaction log analysis—“the study of electronically recorded interactions between online information retrieval systems and the persons who search for the information” (Peters, 1993, p. 41)—and its subset, search log analysis (Jansen, Spink, & Taksa, 2008), provide a means for unobtrusive capturing and analyzing user search queries in various information systems, and empirical data to inform information system design. However, the potential of transaction log analysis has not been used to its full capacity to benefit large-scale domain-specific and domain-independent digital libraries’ development: although several studies (e.g., Khoo et al., 2008; Pan, 2003; Verberne et al., 2010; Zavalina, 2007; 2012) have analyzed transaction logs of individual large-scale digital libraries—American Memory, The European Library, NSDL, IMLSDCC or OH—only two of them undertook content analysis of search queries; user interactions with different types of digital libraries aimed at different user communities have not been compared.

Methods

Domain analysis approach to information science research suggests comparative empirical studies of users in different fields as one of the important ways to study domains [20]. The study presented in this paper examines the following research questions:

- What are the domain-based differences and similarities between user searching in domain-specific large-scale digital libraries with different subject scope and audience?
- What are the differences and similarities between user searching in domain-specific and domain-independent large-scale digital libraries?

The study compared user search queries in transaction logs of three large-scale digital libraries—one domain-independent (IMLSDCC) and two domain-specific belonging to two different domains of US history and STEM education (OH and NSDL)—recorded over a period of 1 year by Google Analytics. For each of the three digital libraries, search queries were grouped with identical queries. This resulted in a total of over 17,000 unique search queries. The following measures were assessed and compared in the query-level analysis of unique search queries:

- distribution of search options: basic and advanced search
- distribution of search query frequencies and search query lengths

a query that is different from all other queries in the search log, regardless of the searcher; all identical queries are usually collapsed together to give the unique queries
6 measured as the number of times the identical search query occurs in the search log sample
7 measured as the number of words per search query
• central tendency measures (mean and median) of search query frequencies and search query lengths.
• frequency distributions of search categories\(^8\) based on the entities of FRBR and FRAD bibliographic models (IFLA, 2008; 2009): class of persons, concept, corporate body, ethnic group, event, family, object, person, place, and work.

**Findings and Discussion**

Overall, a higher proportion of NSDL search queries (39%) used advanced search options and approaches compared to IMLSDCC (20%) and OH (15%). For example, search limits were used in over 38% of NSDL search queries but in only in 13% of IMLSDCC and slightly over 12% of OH queries. Fielded searching, where the user specifies in which metadata field(s) (e.g., author, subject, etc.) the query match should occur, was not used, due to the absence of this option, in domain-specific NSDL occurred very infrequently in domain-specific OH (0.61% of queries) but much more often in domain-independent IMLSDCC (5%). Phrase searching, where the user puts a query or its part in quotation marks, was applied seldom in all three digital libraries, but more often in domain-independent IMLSDCC (1.75 % of search queries) than in two domain-specific digital libraries (1.32% in OH and 0.79% in NSDL).

At the time of analysis, both IMLSDCC and OH were hosted on the same server and, with exception of a color scheme, had identical user interfaces. Thus, any differences in use of search options between these two digital libraries cannot be attributed to the differences in search interface. However, it is possible that interface design may have impacted differences in results between these two digital libraries and NSDL.

Search queries tended to be longer in domain-specific digital libraries (an average of 2.66 words in NSDL, and 2.32 in OH) than in domain-independent IMLSDCC (1.94 words). The median query length however was the same for all three digital libraries: 2 words per query.

The domain-specific digital library with STEM focus substantially differed from two other digital libraries in search query frequency. The average frequency was the distant first in NSDL (6.54), followed by OH (1.91) and IMLSDCC (1.60). The median search query frequency was also substantially higher in NSDL (2) and lower (1) in both OH and IMLSDCC.

\[\text{Figure 1. Distribution of search categories}\]

\(^8\) Intercoder reliability test was performed on 55% of unique search queries in the sample independently coded by the author and another coder; high intercoder reliability (98.44 % or Cohen’s Kappa of .936) was observed.
As shown on Figure 1, there were more similarities in search category distribution between domain-specific OH and domain-independent IMLSDCC than between two domain-specific digital libraries. This, at least in part, can be explained by overlap in content of IMLSDCC and OH which likely causes some overlap in the user base. However, in all three digital libraries the top two search categories were FRBR Group 3, subject, entities: concept (64% of queries) and object (41%) in NSDL, place (34% in OH and 32% in IMLSDCC) and object (31% in both OH and IMLSDCC) in two other digital libraries. Concept was the 3rd most often occurring search category in IMLSDCC (27%) and the 4th in OH (17%). However, the 4th FRBR Group 3 subject entity—event—was observed much less than the other three (10% in OH, 7% in IMLSDCC and even less (7% and 3% respectively) in NSDL. The work search category displayed the most similar frequencies across the three digital libraries; it was observed in 9% of OH, and 8% of IMLSDCC and NSDL queries. The class of persons search category occurred in 8% of OH queries, 5% on IMLSDCC queries, and 3% of NSDL queries. The ethnic group search category occurred in 5% of OH queries but substantially less often in the other two digital libraries: 2% in IMLSDCC and 0.7% in NSDL. Finally, family search category was observed very infrequently in all three digital libraries: in only 0.4% of OH queries, 0.3% of IMLS search queries, and in none of NSDL queries.

Conclusion

Preliminary findings of this exploratory comparative study of user searching in three US-based large-scale digital libraries—two domain-specific and one domain-independent—reveal some similarities as well as several notable differences. Regardless of domain, in all three digital libraries users often search by object and rarely initiate a phrase search. Overall, levels of use of advanced search options vary and cannot be safely attributed to either a specific domain or domain-independence. Distribution of search categories overall differs more drastically between large-scale digital libraries representing different domains—US history and STEM education—than between domain-specific and domain-independent digital libraries. Domain-specific library search queries exhibit longer query lengths and higher query frequencies than domain-independent library queries, although domain-based difference is statistically significant.

Interface design may have contributed to user searching differences between IMLSDCC and OH, which have very similar user interfaces, and NSDL. Additional investigation into this factor is needed and will be carried out by the authors of this poster.

These preliminary results suggest a hypothesis that domain-based differences in user searching are more substantial than the differences between user searching in domain-specific and domain-independent large-scale digital libraries. Studies that will test this hypothesis will need to extend the set of targets to include several domain-independent and a variety of domain-specific digital libraries representing different domains.

References


Abstract

Through this poster we introduce the recent appearance of Neighborhood Book Exchanges (NooX) across North America and outline our plans for an empirical investigation of the phenomenon. Our investigation draws upon social networking theory, information seeking behavior scholarship and literature from community informatics to frame and investigate the following questions: Does the use of NooX engender stronger community ties amongst its users; In what ways do NooX support (or not) users’ information behaviors? Are there identifiable patterns between the NooX content and its location over time? Through this submission we develop a working description of a NooX, propose an initial investigation of the NooX phenomenon, and anticipate contributions to contemporary information theory.

Keywords: neighborhood book exchange, social network theory, social capital, community informatics, information seeking behavior

Introduction

Whether called Book Trading Posts; Little Free Libraries; Pop-up Libraries, or Community Book Exchanges, hand-crafted hutches filled with books are sprouting up in neighborhoods across North America. Over the past two years media articles have covered story after story of Neighborhood Book Exchanges (NooX) taking root in neighborhoods. Some of the media coverage frame NooX as supporting literacy and community building (Gessner, 2011; Kelley, 2011; Jones, 2012; Christian, 2012; Newcomer, 2012), similar to the benefits purported by public libraries. However, those who take an item from a NooX do not need to concern themselves with library cards, Dewey decimal numbers or due dates (Smith, 2012). In addition, NooX cater to walk up (rather than drive and park) visitors, providing a “literary water cooler” atmosphere where neighbors can gather and share their love of reading (Todd Bols in Kelley, 2011).

Reporters’ interviews with the stewards and visitors help detail what the NooX are and how they serve their neighborhoods: they welcome all kinds of visitors and are available for use twenty-four hours a day, three hundred and sixty-five days a year; they brave rainy and snowy weather (Ward, 2011), and receive midnight visitors who browse by flashlight (Scrivener, 2011). The book exchanges rely on an honor system: if you take a book, then you will replace a book and help maintain the resource (Bicak, 2012).

The media hype concerning Neighborhood Book Exchanges is impressive. NooX are credited with bringing people together (Jones, 2012; Gold, 2012; Lee, 2011). One media account describes long-time neighbors meeting for the first time while browsing a NooX (O’Connor, 2012). They are also touted as a convening spot for block parties (Scrivener, 2011; Christian, 2012). In one neighborhood, the visitors became so familiar with their neighbors’ reading tastes that they include notes recommending certain reads for particular people (Atkinson, 2012). Some suggest NooX represent a reaction to the rise of electronic books (Jones, 2012), the weight of information overload, the cuts to government funding for education and libraries, and the loss of community life (e.g., Jones, 2012; Smith, 2012). None of the media articles question the hype concerning the book exchanges’ constructive value.
Despite the recent attention NooX have garnered, to date there are no scholarly investigations into the phenomenon. A rigorous, empirically informed investigation of NooX is needed to explore the motivating factors in their creation and use, the trends in their content and the casual claims regarding their relationship to community building.

**Describing the Neighborhood Book Exchange**

Our working definition of a Neighborhood Book Exchange (NooX) is based on information from the media articles referenced above, visits to numerous NooX in the Greater Vancouver area and informal conversations with a number of NooX stewards when we first became intrigued by the phenomenon. We use the term NooX to describe a small physical container used to host a book exchange system. Although each NooX is unique, they are typically hand-crafted in the likeness of miniature houses. They are enclosed structures with one to many shelves that store a small collection of books and, occasionally, other media. NooX enclosures are typically placed a few feet off the ground on the periphery of a community member’s private property where it is accessible to a public right of way. A volunteer steward builds and/or maintains the NooX. Typically this steward constructs and maintains the NooX physical structure, as well as stocking the initial collection and pruning the content.

Visitors to a NooX include anyone who browses the contents of the NooX, regardless of the frequency of their visits or whether they take part in the exchange. Visitors are not limited in definition to neighborhood residents where a NooX is located. They may take one or many books, leave one or many books, do both, or neither.

Our use of the term exchange in Neighborhood Book Exchange is asynchronous. Changes in content—books taken or books left—happen at different times and transactions may occur between periods of inaction. The exchanges are usually anonymous; the book donations and pickups typically lack attribution.

NooX propagate through both unofficial channels and professionally organized efforts. Some are initiated independently by stewards who have seen or heard of NooX. Other stewards seek financial, logistical and/or technical support from professionally organized efforts, like the Little Free Library organization ([www.littlefreelibrary.org](http://www.littlefreelibrary.org)). The location and time of the first appearance of a NooX—a book exchange that meets our criteria above—is unknown to us.

**Investigating Neighborhood Book Exchanges**

Drawing upon social networking theory, information seeking theory, and community informatics, this project aims to explore the following set of research questions:

- Does the use of NooX engender stronger community ties amongst its users?
- In what ways do NooX support (or not) users’ information behaviors?
- Are there identifiable patterns in the content of a NooX over time with respect to location?

In order to explore the motivations for using NooX and the implications for their communities, we are considering a mixed-methods approach. We plan on administering surveys and conducting face-to-face interviews with individuals who live in NooX neighborhoods (both NooX users and nonusers) along with a longitudinal content analysis of NooX holdings. Because we are in the process of designing our study, we believe this is the opportune time to hone our research questions and methodological approach through engaging experts in social networking, information behavior and community informatics (i.e., iSchool conference attendees).

**Abbreviated Literature Review**

**Community Informatics**

To better understand how an information resource could impact a community, we look to Community Informatics, a field that focuses on the relationships between Information, Communication Technologies (ICT) and community life (e.g., Gurstein (2007); Loader (2010); Stoecker (2005)). We
consider NooX to be a non-digital ICT and a place-based engagement; therefore, this exploration will contribute to discussions in Community Informatics, which focuses almost exclusively on digital ICTs.

Social Capital & Social Networks

Our interest in Social Capital and Social Networking theory (e.g., Granovetter (1973); Williams & Durrance (2008)) is best illustrated by the following questions: Are weak ties formed between actors through a NooX even if these actors never physically meet? Does the presence of a NooX favor bonding or bridging social capital or neither?

Information Seeking Behavior

NooX provide a unique instance to study information seeking behavior. They are peculiar information resources: there are never many items available at one time; new items are added on an irregular basis; the type, topic, and quality is unpredictable; and the collections are disorganized and untraceable. Yet visitors use the NooX as an information resource and, even, make return visits. This information seeking behavior is somewhat striking; why return to a resource that seems so unlikely to provide a book that is relevant or interesting? Three areas within information seeking behavior may provide useful leads: casual-leisure information seeking behavior, browsing behavior, and serendipitous information encountering (e.g., Elsweiler, Wilson, & Lunn (2011); Erdelez (1995); Twidale, Nichols, & Paice (1997)).

Conclusion

This poster will include an analysis of recent media coverage of NooX, our working definition of NooX, highlights from our literature review, an overview of the study design and updates on our progress. Our investigation will be a first step in developing a deeper understanding of the NooX phenomenon. We anticipate contributions to discussions in Social Networking, Social Capital, Community Informatics and Information Behavior research as we investigate non-digital information exchanges through NooX.

References


Granovetter, M. S. (1973). The Strength of Weak Ties. American Journal of Sociology, 78(6), 1360–1380.


Placing the Walls: 3-D Modeling of Chicana/o Murals in LA

David Kim
University of California, Los Angeles
David.js.kim@ucla.edu

Abstract

Placing the Walls is a 3-D simulation model of the Estrada Courts Public Housing, located in East Los Angeles and known for its more than 60 murals that narrate the events and ethos of the Chicana/o Movement of the 70’s. It considers 3-D modeling and simulation as one of emerging digital methodologies for exploring site-specificity and spatial context in the field of cultural informatics and archival studies. The discussion will focus on the technical aspects of building 3-D models with open source applications, as well as the theoretical implications of ‘simulation-as-documentation.’

Keywords: 3-D simulation, cultural informatics, digital humanities, archival studies

Introduction

Developed during the NEH 2011 Vectors Summer Institute on Digital Approaches to American Studies at the University of Southern California, Placing the Walls is a methodological exploration of 3-D modeling and simulation for the analysis of site-specific cultural production in relation to identity and community formations. In the field of classics and archeology, digital models of various ancient sites have been one of the most exciting developments in recent years. In other fields, the ‘games’ environment of digital simulation has been used in classrooms to add the ‘experiential’ and the interactive layer to one’s understanding of various events. While the claims of realism, both objective and ‘augmented,’ is certainly alluring, digital models can be perhaps more productively framed as “representational methodology,” as Chris Johanson argues (2009), for which the models serve as “tinker toys” to test hypothesis and to propose various interpretations. For cultural informatics, 3-D models and simulations are useful for questions and challenges related to how to represent discrete data in some form of continuity.

A collaborative project between Michael Rocchio, a Ph.D. candidate in architecture, and David Kim, a Ph.D. candidate in information studies, Placing the Walls explores how 3-D models can serve both interactive and archival function. Taking the Estrada Courts Public Housing in East Los Angeles as a case study, the model is a digital representation of the site that is famous for more than sixty community murals that were organized as a part of the broader Chicana/o arts movement and LA’s beautification efforts in the 70’s and the 80’s. The model captures all the murals that are currently in existence at the site, including those that are seldom discussed, older images of some of the murals that document their gradual decay, and related archival records, as well as the graffiti and tags that have accumulated over the years. From our digital recreation, we propose spatially oriented analysis of the site, focusing on the contrast between the murals on the exterior walls and those in the interior alleyways and the significance of graffiti in the documentation of the murals.

For the documentation and archival concerns in cultural informatics, 3-D models offer the opportunity to capture spatial contexts on various subject matters, in a manner perhaps more dynamic than, say, a book catalog. By digitally reconstructing a physical site, digital models reveal the constructed-ness of all social spaces, from which one must decide which elements of the site matter and should be captured in the model in order generate productive interpretations of the site. Thus, the digital model’s ultimate aim is not objective realism but analytically useful representation, and its representational advantage is the capacity to gather various pieces information of a site and present them as an interactive ‘whole.’

Copyright is held by the author/owner(s). [this section all Arial 8 pt., no extra/automatic line spacing]
The project is built in combination of various digital platforms available for free to the public. The model itself is generated in Google SketchUp, using high-resolution images we captured at the site and the dimensions we obtained from official records. The KMZ file for the model is then geo-referenced and exported to Google Earth. HyperCities, a geo-temporal authoring platform designed by UCLA, provides the narrative frame for the interpretation of the site, as well as the interface to include related archival materials and descriptions of the murals.

![Figure 1](image.jpg)

*Figure 1.* An image from the 3-D model in HyperCities.

**Reference**

Measuring Serendipity in the Lab: The Effects of Priming and Monitoring

Toine Bogers  
Royal School of Library & Information Science  
Copenhagen, Denmark  
tb@iva.dk

Rune Rosenborg Rasmussen  
Danish National Art Library  
Copenhagen, Denmark  
runras@kunstbib.dk

Louis Sebastian Bo Jensen  
Royal School of Library & Information Science  
Copenhagen, Denmark  
louissbj@gmail.com

Abstract

While the phenomenon of serendipity has proven to be a popular research topic, the issue of how to measure it effectively is still relatively unexplored. We present an exploratory study that contributes to our understanding of this issue by examining the effect of (1) priming people about the concept of serendipity and (2) monitoring participants on how they experience serendipity when searching for information in a controlled environment. Our experiments indicate that it is best to keep such controlled experiments as natural as possible: priming participants about serendipity and monitoring them during their experiments seem to have a negative influence on experiencing serendipity, as they are more likely to induce participants to stay on task instead of exhibiting divergent information behavior.

Keywords: Serendipity, measurement, priming, monitoring, experimentation

Introduction

Serendipity has traditionally been studied in naturalistic settings using interviews (Makri & Blandford, 2012; McCay-Peet & Toms, 2011a), yet systems designed to promote serendipity can often benefit from evaluation in more controlled laboratory settings as well. However, little research has been devoted to recording, evaluating, and measuring serendipity in these more controlled settings. The work described in this paper contributes to measuring serendipity in controlled environments by examining the effect of two different factors on experiencing serendipity when performing information seeking tasks: (1) does priming people about the concept of serendipity cause them to experience more of it, and (2) does monitoring the participants during their information seeking influence how much serendipity they experience? Our goal here is to determine whether we can create certain conditions in the lab that can induce more serendipity, thereby making it easier to evaluate serendipity-enhancing systems in such controlled settings. This could be of great benefit, as serendipity has been shown to be difficult to induce, both in controlled (Erdelez, 2004) and natural environments (Sawaizumi, Katai, Kawakami, & Shiose, 2007).

Related work

The first exploration of inducing and measuring serendipity in a controlled laboratory environment was by Erdelez (Erdelez, 2004), who framed serendipity as information encountering. Erdelez distinguished between foreground and background information needs when encountering information, with the background needs inducing serendipitous encounters when working on an unrelated foreground task. Through a shared course assignment, she fixed the background need for all student participants in her study and inserted information relevant for this background need into the results lists generated by the search engine used in her study. Students were then asked to use this search engine to perform the

Copyright is held by the author/owner(s).
same search tasks. Erdelez found that most students stayed on track and completed the task, with only a handful of students recognizing the serendipity of the planted information.

A more recent foray into the measurement of serendipity is the work by McCay-Peet and Toms (McCay-Peet & Toms, 2011a, 2011b), who developed a measurement instrument for serendipity based on Björneborn’s 10 dimensions of serendipity in physical libraries (Björneborn, 2008). They found that the most important indicators for experiencing serendipity were related to enabling connections between objects, introducing variety and unexpectedness, and inducing curiosity.

Serendipity in the Lab

While there are many factors that could influence how participants in a controlled setting experience serendipity, we focus on two particular ones: priming and monitoring. Priming is the cognitive effect in which exposure to a particular stimulus influences the response to a later stimulus (Reisberg, 1997). Our question here is whether informing participants that serendipity is a part of the experiment makes them more or less likely to experience it. While one could argue for both a positive and negative effect of priming on serendipity, we hypothesize that priming has a negative effect on the number of serendipitous occurrences, because it makes the setting less natural and makes them more aware of the possibility of being distracted.

It is important to keep the laboratory environment as natural as possible in controlled experiments and to remove all possible distractions (Lazar, Feng, & Hochheiser, 2010). One such distraction is formed by the presence of the experimenter. We therefore expect monitoring the participant to have a negative effect on the occurrence of serendipity. In general, we expected our unmonitored, unprimed participants to experience the most serendipity.

Methodology

Participants

We recruited 20 current and former LIS students for our study. All participants rated their Internet experience as being at least ‘average’, with 85% (N=17) rating it as ‘good’ to ‘very good’.

Design of the study

We were interested in studying the influence of two independent variables. The priming variable had two conditions: ‘Primed’ and ‘Not primed’. In the ‘Primed’ (P) condition participants were introduced to the concept of serendipity before the start of the experiment and asked to be aware of it during their participation. They were not informed that serendipity was the main focus of the experiments. Participants in the ‘Not primed’ (~P) condition were not informed that serendipity was a part of the experiments at all. The monitoring variable also had two conditions. Participants in the ‘Monitored’ (M) condition had the experimenter present in the room at all times. After being introduced to the experiment, the participants in the ‘Not monitored’ (~M) condition were left alone during the information seeking part of each task. The experimenter would then re-enter the room to supervise the relevance/interest judging process. Both variables were tested in a between-subjects design with 5 participants randomly assigned to each of the four possible combinations of the two variables.

Experimental procedure

In the introductory phase of the study—after completing a pre-task demographics questionnaire—participants were informed that they had to complete three search tasks using the native search functionalities of one of two websites, Amazon.com and Digg.com, which were selected because of their expected potential for serendipity. Participants had 12 minutes to perform each task, during which they were required to bookmark relevant and/or interesting articles or products. After each task, participants had 3 minutes to rate the bookmarked pages on these two dimensions—relevance to the work task and personal interest—using a four-point graded scale, similar to André et al. (André, Schrafel, Teevan, & Dumais, 2009). On this scale ‘0’ corresponded to ‘not relevant/interesting’ and ‘3’ to ‘highly relevant/interesting’. After this main phase, they were asked to complete a post-questionnaire about their
previous experience with the two websites and the degree to which they experienced serendipity during their searching.

After each completed experiment the search history and cache of the browser on the computer used in the experiments were cleared. We used Morae\(^1\) to record the interactions of the participants with the computer. Participants were offered to receive a list of the pages they had bookmarked during their sessions by e-mail.

During the main phase, participants were asked to complete two fixed work tasks and come up with a third task representing their own interests, but they were kept unaware that the true focus of the experiment was on serendipity. Task ordering was randomized. Our two cover tasks were chosen to represent both an informational and a transactional information need\(^2\) (Broder, 2002). In creating our work tasks, we followed Borlund’s recommendations for having participants perform simulated work tasks (SWT) with realistic cover stories, and using graded, multi-dimensional relevance judgments (Borlund, 2000). The cover stories for our informational and transactional SWTs were the following:

In May 2011, Microsoft acquired Skype, whose software supports free voice and video calling over the Internet. As a dedicated Apple user you are worried about whether this will affect your weekly Skype sessions with your sister living in Brazil. Use Digg.com to find out more about Microsoft’s plans for keeping Skype free-of-charge and the future compatibility with Apple products.

Your best friend is turning 30 and you would like to get him/her a very unique gift of up to $100. You know your friend is very passionate about rock music. Try to find some unique collector’s items or rock memorabilia on Amazon.com for the stated $100 budget.

In line with Borlund (Borlund, 2000), we also asked participants to come up with a task representing their own interests. To prevent them from searching for relevant information beforehand, participants were asked during the experiment to come up with an information seeking task of personal interest. The task could either be of an informational or a transactional nature, and had to be carried out using Digg.com or Amazon.com respectively. If a user could not come up with a personal task, they could select one of two prepared backup tasks, so that all users completed three different tasks.

Results

Participants were asked to judge bookmarked pages on both relevance and interest, the latter dimension corresponding to serendipity. We define a bookmarked page (henceforth: hit) as serendipitous if a participant judged it as non-relevant (relevance grade = 0) but interesting (interest grade ≥ 2). Combined, our 20 participants found 57 serendipitous hits in total, for an average of 2.85 serendipitous hits per participant.

When we look at the effect priming has, we find some indications that not priming participants resulted in more serendipitous hits: primed participants had an average of 2.7 hits, while unprimed users had an average of 3.0 hits (N=10 in both cases). However, this difference is not significant (t(18)=0.31, p = 0.757). We did find a significant difference in the number of opened Web pages, with unprimed participants opening more pages (t(18)=1.86, p = 0.036), suggesting that priming participants almost acts like a warning to stay on task.

There are indications that not monitoring participants resulted in more serendipity: monitored participants found 2.6 serendipitous hits on average, unmonitored participants 3.1 hits on average (N=10 in both cases). These differences were not significant (t(18)=0.53, p = 0.605).

Looking at the number of serendipitous hits for the four different conditions, we see the following distribution: participants in the P/M condition found 2.6 hits on average, P/¬M participants found 2.7 hits, ¬P/M 2.6 hits, and ¬P/¬M 3.4 hits (N=5 in all cases). While we do see an indication that our unprimed,

\(^1\) Morae is usability testing software that can capture audio, on-screen activity, and keyboard/mouse input.

\(^2\) We did not consider navigational needs as we did not expect them to lead to much serendipity because of their shorter and more direct nature.
unmonitored participants experienced the most serendipity—in line with our hypothesis—an ANOVA test revealed no significant differences between these four conditions (F(3,16) = 0.14, p = 0.933).

We found no significant differences in the number of serendipitous hits for personal tasks vs. assigned tasks. Previous experience with the websites used also did not have a significant influence on the number of serendipitous hits (F(1,20) = 1.55, p = 0.235).

Finally, an analysis of the Morae recordings and the timestamps of bookmarked serendipitous hits gave some insight into what people were doing when they experienced serendipity: in 42 cases participants were searching for information, in 13 cases they were browsing, and in two cases they experienced serendipity already on the frontpage of the websites.

Discussion & Conclusions

In this paper we have presented the results of an exploratory study on the effect of priming and monitoring on experiencing serendipity during information seeking. Our experiments indicate that it is beneficial to keep controlled experiments designed to measure serendipity as natural as possible: priming participants about serendipity and monitoring them during their experiments seems to have a negative influence on experiencing serendipity. A possible explanation for this could be that our participants consisted exclusively of LIS students, who prided themselves on their searching abilities and perhaps stayed on track more than other participants would have. Counter to expectations, tasks of personal interest to the participant did not seem to induce significantly more serendipitous occurrences.

Future work would require a more critical look at our operational definition of serendipity—interesting but non-relevant results—which is likely to be an oversimplification of the concepts. Using a more sophisticated measurement instrument as the one developed by McCay-Peet and Toms (McCay-Peet & Toms, 2011a, 2011b) would be a promising improvement. Increasing the size of the study with more variation in participant backgrounds would also be a necessary step in producing more realistic results.

References


The Role of Digital Libraries as Boundary Objects Within and Across Communities

Adam Worrall
Florida State University
College of Communication and Information
apw06@my.fsu.edu

Abstract

Despite increased study of social contexts within information science, it is still unclear if and how digital libraries support and facilitate collaboration, communities, and other social contexts. This poster presents a study that will examine the role of the LibraryThing and Goodreads digital libraries, as social phenomena and boundary objects, in information behaviors and activities taking place within, between, and across multiple existing and emergent communities. The study will focus on the two key phenomena of communities and collaboration, under a theoretical framework drawing from Star’s boundary object theory, Strauss’s social worlds perspective, and Burnett and Jaeger’s theory of information worlds. Data will be collected from the two cases using a sequential, multi-phased mixed methods design employing content analysis, a survey, and interviews. The study should have significant implications for digital library research and practice and for related research on social networking, social media, and social Web services.

Keywords: digital libraries, boundary objects, communities, social informatics, social contexts

Introduction and Background

Despite an expressed need—as far back as Bush (1945)—for social contexts of information to be considered under a social paradigm, many early information retrieval systems focused on technology (see e.g. Raber, 2003; Smith, 1991). This was no less true of early digital library (DL) research and practice, which showed echoes (Borgman, 1999) of the same paradigmatic unrest present in information science as a whole. The rise of social informatics (Kling, 1999; Sawyer & Tapia, 2007) and social constructionism (Tuominen & Savolainen, 1997; Weinberg, 2009), as part of a social paradigm for information science, have led to greater understanding of the importance of contextual views of information, ICTs, and information behavior (Courtright, 2008), including DLs (e.g. Lynch, 2005; Van House, 2003). Nevertheless, no one approach has been determined to be the way to support the social contexts of DLs, and further research grounded in existing literature, theory, and practice is necessary into if and how DLs facilitate collaboration, communities, and interaction in social contexts. This poster presents a study aiming to address this by examining the role of two DLs, as social phenomena and boundary objects, in information behaviors and activities taking place within, between, and across multiple communities.

Drawing from Borgman (1999) and other literature, a social digital library can be defined as

- having one or more collections of digital content collected on behalf of a user community;
- offering services, relating to the content, by or through the DL to the user community; and
- being one or more—or part of one or more—formal or informal organizations managing these content and services.

Acknowledgements: I thank Michelle M. Kazmer, Gary Burnett, Sanghee Oh, Deborah J. Armstrong, and the anonymous reviewers for their comments and feedback on this research in progress.


Copyright is held by the author/owner.
All of these should be considered in light of various contexts, especially the social contexts. Viewing DLs as social parallels the roles of physical libraries (Pomerantz & Marchionini, 2007, p. 506), which are not just physical collections and technical services but physical and conceptual spaces “link[ing] people to ideas and to each other.” Since a traditional role of physical library environments is to serve as inherently social spaces, DLs should improve their support for collaborative information behaviors, lest social opportunities to seek, use, and share information become diminished or lost as libraries become increasingly digital and hybrid in nature.

The phenomena of communities and collaboration are key elements of this problem. A user community may consist of smaller communities or groups, adopting the subcultural view pioneered by Fischer (1975) and incorporating flexible use of conceptions of community used in calls for social digital libraries (e.g. Gazan, 2008; Van House, 2003) and related research areas (see e.g. Ellis, Oldridge, & Vasconcelos, 2004; Haythornthwaite, 2007; Veinot & Williams, 2012). The DL collects content for use by these communities, offers services to them, and is associated with organization(s) that are a form of community. A major objective of DLs is to support, construct, and build these differing kinds of “knowledge communities” that use their content and services (Bearman, 2007, p. 245). DLs can and should support communities’ “internal workings … and their links to the rest of the world” (Agre, 2003, p. 227; see also Star, Bowker, & Neumann, 2003). This requires supporting social contexts: collaboration within and across communities, the building of existing and emergence of new communities, and individual and collaborative information behavior. In most cases, such collaboration centers around a common overall project, goal, interest, or practice (Gunawardena, Weber, & Agosto, 2010), but collaboration may also be serendipitous (see Erdelez, 2005; Foster & Ford, 2003; Talja, 2002).

Framework and Approach

This study conceives of DLs as socially constructed phenomena, using the perspectives offered by the social paradigm, social constructionism, and social informatics. The study draws on a theoretical framework including Star’s (1989; Star & Griesemer, 1989) boundary object theory, Strauss’s social worlds perspective (Clarke & Star, 2008; Strauss, 1978), and Burnett and Jaeger’s (2008; Jaeger & Burnett, 2010) theory of information worlds. Under this framework, because they are used by and cross the boundaries of multiple social worlds, information worlds, and communities, social digital libraries should act as socially constructed boundary objects. They should

- adapt to the “local needs” of as many of these worlds and communities as possible (Star, 1989, p. 46);
- reconcile and translate “meanings” and understandings across these worlds to allow users to “work together,” collaborate, and interact (Star & Griesemer, 1989, pp. 388–389);
- support the emergence of localized and common social norms, social types, information values, and information behaviors shared—to varying and overlapping extents—by the different information worlds using them (Burnett & Jaeger, 2008; Jaeger & Burnett, 2010);
- act as common sites and technologies for users to engage in information-based activities (Strauss, 1978), including collaboration and information sharing; and
- support the possible emergence of broader communities, social worlds, and information worlds as they converge, coalesce, and reconcile portions of the multiple communities they serve.

Research Design and Methods

This study will focus on two cases, LibraryThing and Goodreads, which are digital libraries and Web sites for readers and lovers of books. LibraryThing and Goodreads feature digital content—from outside organizations and users—collected for their users and user communities, services relating to the content and for their user communities, and formal and informal organizations managing the content and services; as such they are social digital libraries. The study will answer two research questions:

1. What role(s) do LibraryThing and Goodreads play, as boundary objects, in translation and coherence between the existing social and information worlds they are used within?
2. What role(s) do LibraryThing and Goodreads play, as boundary objects, in coherence and convergence of new social and information worlds around their use?
A case study approach (Yin, 2003) will employ a mixed methods research design, using qualitative and quantitative methods together to combine their strengths, minimize their weaknesses, and obtain a fuller understanding of social digital libraries. The research design is a variation on Creswell and Plano Clark’s (2011) multiphase design incorporating elements of their sequential designs. Qualitative and quantitative data will be collected and integrated in sequence; qualitative data is prioritized, but not at the expense of quantitative data collection; and multiple methods are used within the one study. The study will incorporate the viewpoints of multiple perspectives, a strong theoretical framework, and a multi-leveled approach to analysis.

Three phases of data collection are planned. First, a content analysis phase will collect and analyze messages from LibraryThing and Goodreads’ group discussion boards. Systematic random sampling will be employed to collect about 500 messages across 10 of these groups, taken from the groups most active and popular in the days prior to data collection. Messages will be collected by accessing the DLs’ group discussion boards and saving individual threads. These will then be coded and analyzed using key concepts and phenomena from the theoretical framework. Preliminary findings from this phase are expected to be available by the time of the iConference.

Second, a survey phase will use an online questionnaire to obtain data from users of LibraryThing and Goodreads. Invitations to participate will be sent to LibraryThing users who posted messages analyzed in the previous phase; invitations will also be posted in the ten LibraryThing and Goodreads groups selected. The survey will include Likert scaled questions on the concepts used in the theoretical framework and demographic and usage questions. Two reminders will be sent during the survey collection process, which will last six weeks and should obtain at least 300 responses. Participants will be entered into a drawing for 10 $25 Amazon.com, Barnes and Noble, or Books-A-Million gift cards as compensation. Appropriate descriptive and inferential statistical methods will be used to confirm the reliability and validity of the scales and analyze the results.

Third, a phase of semi-structured qualitative interviews will identify users for whom follow-up interviews could lead to insightful data. At least 15 users across the two DLs should be interviewed, but the final number may vary depending on when saturation is reached. The semi-structured interviews will follow pre-planned questions and themes drawn from the theoretical framework, but additional follow-up questions, probes, and prompts may emerge from the conversation. Critical incidents (Flanagan, 1954) of times when users interacted with others using the LibraryThing or Goodreads DLs should provide a rich environment and context within which to explore these themes. Interviews will take place using online audiovisual media or telephone and will be audio recorded using computer software. Interviews will later be transcribed, then—as with messages—coded and analyzed using key concepts and phenomena from the theoretical framework.

The field of digital library research, and by extension the information science field, will benefit from a fuller understanding of the role and uses of social digital libraries within and across worlds and communities, as should come from this study. Significant implications are expected to be identified for DL design, usability, and development; provision of services in and by DLs; and use of DLs by users and user communities. The study will further benefit related research on social networking, social media, and social Web services.

References


Automatic Detection of Search Tactic in Individual Information Seeking: A Hidden Markov Model Approach

Shuguang Han
University of Pittsburgh
shh69@pitt.edu

Zhen Yue
University of Pittsburgh
zhy18@pitt.edu

Daqing He
University of Pittsburgh
ah44@pitt.edu

Abstract

Information seeking process is an important topic in information seeking behavior research. Both qualitative and empirical methods have been adopted in analyzing information seeking processes, with major focus on uncovering the latent search tactics behind user behaviors. Most of the existing works require defining search tactics in advance and coding data manually. Among the few works that can recognize search tactics automatically, they missed making sense of those tactics. In this paper, we proposed using an automatic technique, i.e. the Hidden Markov Model (HMM), to explicitly model the search tactics. HMM results show that the identified search tactics of individual information seeking behaviors are consistent with Marchionini’s Information seeking process model. With the advantages of showing the connections between search tactics and search actions and the transitions among search tactics, we argue that HMM is a useful tool to investigate information seeking process, or at least it provides a feasible way to analyze large scale dataset.

Keywords: information seeking process; search tactics; Hidden Markov model

Background & Motivation

Information seeking process (ISP) is one of the major areas in information seeking behavior research. There are several well-known models for describing individual’s information seeking process. Both Kuhlthau’s [1] and Ellis’[2] models present holistic views of information seeking from the initiation stage to the ending stage, whereas Marchionini’s model [3] consists of eight stages and it focuses on describing possible transitions among each of them. In addition to the above-mentioned macro-level investigations of search processes focusing on qualitative constructs of stages and context in ISP, there are also several empirical studies that look into micro-level search actions in ISP.

Some studies examined search processes through the actions in the user logs. For example, Holscher and Strube [4] compared action sequences between Internet experts and newbies. Chen and Cooper [5, 6] used both stochastic model and clustering techniques to examine search tactics in a Web-based library catalog. The problem of those approaches is that they missed explaining user intentions behind user actions. Xie and Joo [7] raised the importance of investigating transitions of search tactics as a means of examining search processes. They manually coded the transaction logs using a predefined scheme of search tactics. Then, a five-order Markov chain was adopted to find the common search tactics in user’s behavior sequence.

Based on the literature review, we can see that investigation on information seeking process in individual user is an active research topic. Particularly, search tactics had been recognized as a mean of investigating search processes. However, most of previous researches either focus on a global qualitative analysis of search stages, or highly rely on manually coding of users logs. The manually coding of user actions is difficult to be expanded or used in a different or large-scale dataset. Although automatic

Acknowledgement: This work was partially supported by the National Science Foundation under Grant No. 0704628 and IIS-1052773.


Copyright is held by the author/owner(s).
methods have been explored, they usually missed showing the connections between search tactics and search actions. The identified search tactics are simply the aggregation of certain user actions while the connection between user action and search tactic is missing. In this paper, by treating the sequence of user behaviors as a Markov chain, we modeled users’ search tactics as hidden variables. Then, a HMM algorithm is used to identify those hidden search tactics. Each search tactic is then represented by a probability distribution over user actions, i.e. the probability of each search tactic generates each user action. Another outcome of HMM is the transition probabilities among the identified tactics, i.e. the probability of each search tactic transfers to another search tactic (including itself).

**Automatic Detection of Search Tactics**

**Model Search Tactics**

The HMM model for search tactic identification is depicted in Figure 1. Suppose we have a sequence of user actions from O1 to ON. We model search tactics as hidden states and assume each action is generated by one search tactic, from T1 to TN. Assume that we have R different types of user actions (from A1 to AR) and M different types of search tactic (from S1 to SM). Each search tactic Ti (1 ≤ i ≤ N) is one of Si (1 ≤ i ≤ M), and each user action Oi (1 ≤ i ≤ N) is one of Ai (1 ≤ i ≤ R). A HMM has several parameters: the number of user actions N, the number of user action types R, the number of search tactics M, the transition probabilities among search tactics, i.e. the probability between Si (1 ≤ i ≤ M) and Sj (1 ≤ j ≤ M), and the emission probabilities from Si (1 ≤ i ≤ M) to Ai (1 ≤ i ≤ R). By only defining the M, the Baum-Welch algorithm [11] could be used for parameter estimation.

![Figure 1: A Hidden Markov Model for Search Tactics](image)

**Model Selection**

It is still an open issue in determining the M in a HMM. A large M will increase the sequence likelihood because more parameters can describe data more precisely, but with a high risk of being over-fitting. On the other hand, a small M may be insufficient to describe the data. Choosing an optimal M is a model selection problem and Bayesian information criterion (BIC) [12] is usually adopted.

In order to use the BIC, we need to calculate the number of parameters NP. Since we have M search tactics and T actions, so there will be M*M transition probabilities and M*T emission probabilities. In HMM, we also need to define the prior probabilities of each search tactic, the number of which is M. Considering the constraint on the summation of probabilities equals to 1, the NP = M*(M-1)+M*(T-1)+(M-1). Suppose the sample size is S and the data likelihood is L. Then, the BIC is defined in Eq. (1). A large log(L) and less parameter NP are preferred. The smaller of BIC, the more preferred the model is.

\[
BIC = -2 \times \log(L) + \log(S) \times NP
\]  
Eq. (1)

**User Study and Dataset**

In order to obtain the real data for users’ information seeking behavior, we conducted a user study using a web search system that we built [8]. Seven students from the University of Pittsburg who use computers on a daily basis were recruited to participate the study. Each participant was required to
complete two search tasks with one on academic topic [9] and other one on leisure topic [10]. The goal of each task is to collect as much relevant information as possible.

Through analyzing the search logs obtained, we identified five typical types of search actions which are Query, View, Save, Workspace, Topic (See Table 1). The user study gives us a real dataset that consists of fourteen different behavior sequences (two tasks × seven participants).

Table 1: User search actions

<table>
<thead>
<tr>
<th>Actions</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query (Q)</td>
<td>A user issues a query or clicks a query from search history</td>
</tr>
<tr>
<td>View (V)</td>
<td>A user click on a result in the returned result list</td>
</tr>
<tr>
<td>Save (S)</td>
<td>A user saves a snippet or bookmarks a webpage</td>
</tr>
<tr>
<td>Workspace (W)</td>
<td>A user clicks, edits or comments on an item saved in the workspace</td>
</tr>
<tr>
<td>Topic (T)</td>
<td>A user clicks on the topic statement for view or leaves comments</td>
</tr>
</tbody>
</table>

Result and Discussion

HMM Results

The BIC evaluation of our HMM model indicated that the optimal number of hidden states is 5 (See Figure 2), so our following analysis have all been based on M=5.

![Figure 2: The BIC Evaluation](image)

The emission probabilities are shown in Table 2. We remove the probabilities that are smaller than 0.05 for better visualization. Table 2 shows association between search tactics and user actions. For example, S1 refers to a search tactic that is mostly represented by Query, S3 refers to Save, S4 refers to Workspace, and S5 refers is represented by both Workspace and Topic. It is also interesting to observe that both S2 and S3 refer to View. To reveal the differences between S2 and S3, we looked at their transition probabilities.

Figure 3 visualizes the transition probabilities between all the search tactics. Each cell denotes a transition probability from the search tactic in the row to the search tactic in the column. The darker the cell is, the bigger the transition probability. The visualization is shown in Figure 3. Now it is clear that S2 tends to transmit to S3 while S3 tends to transmit to S4. It is probably that S2 is associated with quick scanning the results whereas S3 is more carefully checking results that lead to finding and saving relevant documents.
Table 2: Search Tactics and Emission Probability

<table>
<thead>
<tr>
<th></th>
<th>Q</th>
<th>V</th>
<th>S</th>
<th>W</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.92</td>
<td></td>
<td></td>
<td></td>
<td>0.06</td>
</tr>
<tr>
<td>S2</td>
<td></td>
<td>0.97</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td></td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td></td>
<td></td>
<td></td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.67 0.32</td>
</tr>
</tbody>
</table>

Figure 3: Transition of Search Tactics

Comparison with Marchionini’s ISP model

Marchionini’s ISP model is one of the well-known ISP models. It provides a clear description of user search behaviors in digital environment. His model divides user’s information seeking process into eight sub-processes, and provides the transition possibilities among those sub-processes.

Because of the relative similarity between Machionini’s model and our HMM model, we directly compare these two models. Some of Machionini’s sub-processes are mapped jointly to one HMM search tactic. For example, the sub-process “select source” is predefined in our system because we use Google results by default. “formulate query”, as its defined in the model is a cognitive work, which is difficult to be recorded. Therefore, we treat the “select source”, “formulate query” and “execute search” as an integrative process called “query”.

The comparison results are shown in Table 3. The search tactics obtained from HMM model can reasonably been mapped to sub-processes in Marchionini’s ISP model. According to the ISP model, the main transition path is S5 → S1 → (S2 → S3) → S4, which almost perfectly matches to the darkest areas shown in Figure 3. Our HMM model mark the one “examining results” sub-process in the ISP model as two separate search tactics S2 and S3.

Table 3: Mapping from sub-process to HMM patterns

<table>
<thead>
<tr>
<th>Sub-processes</th>
<th>Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Define Problem</td>
<td>S5</td>
</tr>
<tr>
<td>Select Source</td>
<td>S1</td>
</tr>
<tr>
<td>Formulate Query</td>
<td></td>
</tr>
<tr>
<td>Execute Query</td>
<td></td>
</tr>
<tr>
<td>Examine Results</td>
<td>S2, S3</td>
</tr>
<tr>
<td>Extract Information</td>
<td>S4</td>
</tr>
<tr>
<td>Reflect/Iterate/Stop</td>
<td>S5</td>
</tr>
</tbody>
</table>
Conclusion

Search tactics have been recognized as a way of investigating information seeking processes. However, most of previous works are either based on predefined search tactics or simply aggregation of sequential user actions. In this paper, we adopted a Hidden Markov Model (HMM) approach to solve these two problems. By introducing the search tactics as hidden variables HMM can build connections between search tactics and search actions. Our result shows that HMM also produce a reasonable performance in our experiment because the result shows a considerable agreement with Marchionini’s ISP model. HMM is a standard machine learning algorithm, which can easily be applied in a different, or a very large dataset.

References

Abstract

Vector Space Modeling (VSM) and Language Modeling (LM) are the two most influential retrieval models currently. They appear to have different perspectives and use different mathematical tools. However, they are actually closely related. The current study analyzed their relationship, compared their weighting schemes and revealed their connections. Our findings suggest that although the VSM and the LM originated from different perspectives, they are closely related. The backbone of the LM weighting is still a TF-IDF like weighting scheme.

Keywords: vector space modeling, language modeling, relationships

Introduction

Vector Space Modeling (VSM) and Language Modeling (LM) are the most popular information retrieval models currently. They provide different ways to represent documents and queries, as well as different means to evaluate documents against queries. The VSM was first proposed in Salton and McGill (1983). It uses linear algebra tools to model the documents and terms. A document is represented as a vector and the terms are its elements. On the other hand, the LM was first brought to information retrieval by Ponte and Croft (1998). It is a branch of probabilistic models. A document is viewed as a language model, which is essentially a probability distribution over its terms. At the first glance, the two models take quite different perspectives and use very different mathematical tools. However, it has been speculated that the two models are related. A search on the literature suggests few studies have thoroughly examined the relationship between the two models. Zhai and Lafferty (2001) briefly discussed the connections when studying the smoothing in the LM. Robertson (2004) also pointed out that the weighting in the LM achieves a similar effect as the classical TF-IDF weighting in the VSM by somewhat different means. However, no deep analysis and further discussion was provided with respect to how they are related. The purpose of this study is to uncover the connection between the two models and provide an in-depth insight into their relationship. A good understanding of their relationship will help us to better interpret our search results.

Brief Description of the Two Models

Vector Space Modeling

The VSM uses vectors to represent documents and the elements of a vector consist of words appearing in the collection. The mathematical representation is given as follows:

$$V = \begin{bmatrix} v_1 \ldots v_m \\ v_2 \ldots v_m \\ \vdots \vdots \\ v_n \ldots v_m \end{bmatrix}$$
The rows of the matrix are defined as documents in the vector space while the columns of the matrix are defined as the terms which are used to describe or index the documents in the vector space. This matrix is commonly referred to as the document-term matrix. An element \( v_{ij} \) \((1 \leq i \leq n, 1 \leq j \leq m)\) in the document-term matrix reflects the normalized weight of the indexing term \( t_j \) assigned to the document \( d_i \). Here \( n \) and \( m \) are the number of documents and indexing terms in the vector space respectively. The prominent TF-IDF defines the term weight to be proportional to the term frequency in the document and inversely proportional to the number of documents that contain the term (i.e. document frequency):

\[
v_{ij} = tf_{ij} \cdot \log \frac{|C|}{n(t_j)}
\]

where \( tf_{ij} \) is the term frequency weighting of the \( j \)th term in \( i \)th document, \( n(t) \) denotes the document frequency, and \( |C| \) denotes the number of documents in the collection. To control the effect of the document length, the document length normalization is usually applied to the TF component (Harman & Voorhees, 2006).

**Language Modeling**

Language modeling was first used in natural language processing to model the probability of a sequence of words. Ponte and Croft (1998) introduced LM to information retrieval by considering retrieval as a generative process. The LM ranks the documents according to their probabilities to generate the query terms. To estimate the probability of seeing a term in a document language model, the maximum likelihood estimation is usually adopted:

\[
\tilde{P}(t|D) = \frac{f(t,D)}{|D|}
\]

where \( f(t,D) \) is the frequency of a term in the document, and \(|D|\) is the document length. Because a document language model is estimated from a limited sample (i.e. one document), it is likely to have a data sparseness problem. So, smoothing becomes apparent for the LM. A commonly used smoothing method is the Jelinek-Mercer smoothing. It mixes the probability estimated from the document with the one from the collection:

\[
\tilde{P}_{\text{smooth}}(t|D) = \lambda \tilde{P}(t|D) + (1 - \lambda) \tilde{P}(t|C)
\]

where the maximum likelihood estimation of the \( \tilde{P}(t|C) \) is the collection frequency of the term divided by the total term count in the collection.

**Model Comparison**

Both models naively assume the independence of terms in documents. On the surface, the two models provide similar functions (represent documents and weight terms) from quite different perspectives. However, a detailed investigation uncovers their connections. First, a vector in the VSM and a probability distribution in the LM are similar in containing the term weights although they have very different mathematical intuitions (i.e. geometry vs. probability). The only difference is that a probability distribution is normalized to sum to one while a vector does not have such requirement. In terms of their term weighting schemes, The TF-IDF is actually closely related with the probability weighting method in the LM. The standard LM weighting can be decomposed into the following equation (Zhai & Lafferty, 2001):

\[
\log p(Q|D) = \sum_{i:C(q_i|D) > 0} \log \frac{p(q_i|D)}{(1-\lambda) p(q_i|C)} + m \log (1-\lambda) + \sum_{i=1}^{m} \log p(q_i|C)
\]
where \( p(q_i|D) \) denotes the probability of a query term in a document, \( p(q_i|C) \) is the probability of a query term in the collection, \( \lambda \) is a parameter to control the amount of smoothing. In equation 5, the only component that influences the rankings is the first addend, which indicates that the LM weighting is actually proportional to the term frequency in the document and inversely proportional to the collection frequency. This is very close to what is described in the TF-IDF except that the collection frequency is used in the LM instead of the document frequency. Therefore, the relationship between the VSM and the LM is rooted in the relationship between the collection frequency and the document frequency. However, in terms of how the collection frequency is related with the document frequency, the authors did not provide any further evidence. To investigate this relationship, we selected a number of different document collections from TREC, and examined the collection frequencies and the document frequencies of the terms. The following section will report our results.

**Collection Frequency versus Document Frequency**

A number of representative document collections from TREC were selected. This includes the collections for the Genomics Track 2006, the WT10G, the TREC-6, and the Medline collection. The Genomics Track is a collection of full text academic articles in the field of biology linked with genomics information. The WT10G collects a large number of English web pages. The TREC-6 contains newspaper and government records. The Medline collection is a small collection of documents from Medline plus database. For each collection, TREC provides a number of test topics and their relevance judgments. The data collections were indexed by the Indri search engine (www.lemurproject.org). Stop words were removed and stemming was applied. The descriptive statistics of the document collections are listed in Table 1.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Corpus size</th>
<th># of queries</th>
<th>Avg. doc length</th>
<th># of unique tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics</td>
<td>162,259</td>
<td>64</td>
<td>6,595</td>
<td>2,075,859</td>
</tr>
<tr>
<td>WT10G</td>
<td>1,692,096</td>
<td>100</td>
<td>617</td>
<td>5,256,472</td>
</tr>
<tr>
<td>TREC-6</td>
<td>556,077</td>
<td>50</td>
<td>526</td>
<td>767,503</td>
</tr>
<tr>
<td>Medline</td>
<td>1,033</td>
<td>30</td>
<td>155</td>
<td>9,537</td>
</tr>
</tbody>
</table>

To understand the connection between the LM weighting and the TF-IDF, we examined the correlations between the collection frequencies and the document frequencies of the query terms. The query terms are obtained from the title field of the TREC retrieval topics. Those query terms that do not appear in the collection are dropped as they are not affecting the retrieval. Spearman’s \( \rho \) is reported since the data could be highly skewed. The results are provided in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Genomics</th>
<th>WT10G</th>
<th>TREC-6</th>
<th>Medline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman’s ( \rho )</td>
<td>0.966**</td>
<td>0.986**</td>
<td>0.989**</td>
<td>0.952**</td>
</tr>
</tbody>
</table>

**indicates significant correlation at 0.01 level (2-tailed)

From Table 2, we can tell that the collection frequencies of the query terms are both strongly and significantly correlated with their document frequencies in all the collections. This indicates the two measurements are highly accordant in weighting the query terms.
Figure 1 gives the scatter plots of the document frequencies and the collection frequencies in the four collections. All the charts showed the strong positive correlations between the two measurements. Given that the TF-component is exactly the same as the term probability in the document, we can conclude the TF-IDF weighting and the LM weighting are comparable. However, it should be noted that the two weighting are not exactly the same. First, there are some differences in their calculations. For example, TF-IDF applies logarithm on the IDF part before multiplying the TF part, while according to equation 5 the LM does the multiplying first and then apply the logarithm. Second, the strong and significant correlations do not imply the sameness. Some terms can have similar collection frequencies but different document frequencies or vice versa.

Conclusion

The VSM and the LM are the two most influential retrieval models currently. They appear to have different perspectives and use different mathematical tools. However, they are actually closely related. The current study analyzed their relationship, compared their weighting schemes and revealed their connections. The TF component (with document length normalization) in the TF-IDF weighting is exactly same as the probability of seeing a term in a document language model. The IDF component is implicitly related to the smoothing methods in the LM. After decomposing the LM weighting, we have found that the difference is that the LM uses the collection frequency instead of the document frequency in the VSM. An examination on the relationship between the collection frequencies and the document frequencies of the terms in several representative TREC collections indicates that they have both strong and significant correlations. Therefore, we conclude that although the VSM and the LM originated from different perspectives, they are closely related. The backbone of the LM weighting is still a TF-IDF like weighting scheme.
References


Abstract

This poster describes the creation of a multi-modal, interactive digital report as an example of emerging configurations of scholarly communication. The digital report is a web-based artifact including textual, aural, graphic, and video content presented in a loose narrative structure, and is designed to return agency of narrative creation and sense making to the user. The value, creation, and challenges associated with designing and creating the digital report are discussed.

Keywords: scholarly communication, emerging configurations, narrative structure

Introduction

Each August and September, Monterey Bay, California becomes a destination for blue whale watchers. There is something captivating in the whale’s momentary breach, and whale watchers know this. Although only one of the authors of this paper has been out to sea, in our ways, we are both avid whale watchers. Like enormous cetaceans at the ocean’s ceiling, new scholarly practices breach the surface of opaque infrastructure and provide glimpses at the shape of things to come. There is something captivating here, and we know this.

Below the surface of murky infrastructural waters, the production and dissemination of all species of scholarly artifacts is a complicated process. Universities, individual researchers, research teams, publishing houses, libraries, and archives are only a few of players in the knowledge production process. Even in the context of such a reductive list, the constellation of relationships can be considered a complex system, or a collection of “diverse, connected, interdependent entities whose behavior is determined by rules, which may adapt, but need not” (Page, 2011, 6). In his discussion of complex systems, Page states, “the interaction of [interdependent] entities often produces phenomena that are more than the parts” and refers to these phenomena as “emergent” (2011, 6). The idea of emergent phenomena, like whale unexpectedly breaching from the depths of a system, serves as an illustrative analogy: new species of scholarly artifacts occasionally emerge from the complexity of scholarly communication. This poster comprises an entry point into a conversation about the emergence of new types of scholarly artifacts by focusing on a recent digital report: “Emerging configurations of knowledge production: A digital report on an NSF workshop ‘emerging configurations of the virtual and the real’” (Knobel, et al., 2012).

Sustained Artifice: The Textual Archive

Research dissemination techniques stabilize to form the artifice\(^1\) of infrastructure through repetition and gradual modification, “regula[tion] and repress[ion]” (Voss & Werner, 1999, i). Although this process is incremental, the potential of alternate infrastructures reveals itself at discrete moments. With the advent of the printing press and the subsequent construction of scholarly communicative infrastructure, the scholarly community witnessed and artificially sustained a once-emergent communicative means. Such an emergence signaled the form and functionality of what has been called the positivist discourse that comprises the scholarly archive (Foucault, 1972).

\(^1\) Artifice is used in the sense of “The Sciences of the Artificial” (Simon, 1964).


Copyright is held by the author/owner(s).
We know much about the scholarly archive comprised of print materials. We are also adept at communicating through it. But, this is no longer enough. The prevalent conceptualization and structure of artifacts composing the scholarly archive no longer align with the complex multi-modal potentials of scholarly communication. As Voss and Werner express, “the leaves of the archives are again fluttering out, this time into the windless region of hyperspace” (1999, i). It is necessary to investigate alternative forms of scholarly communication, and to study the structure, design, behavior, and use of expressed scholarship that looks less and less like the traditional, purely textual, and easily archivable formats that are the reticent-to-change foundations of scholarship.

The Anatomy of a (Less) Hopeful Whale

While analyzing qualitative data gathered at an NSF-funded workshop entitled, ‘Emerging configurations of the virtual and the real,’ the authors began engaging with exciting new incremental forms of scholarly expression: the database-like interactive digital report. Close engagement with an extant report, Blue Velvet (Goldberg & Hristova, 2007) served as a guide. Blue Velvet is an impressive online interactive archive of materials relating to New Orleans and Hurricane Katrina. Despite its brilliance, Blue Velvet exists as something of a hopeful monster (Law, 1991). Its format is so radically different from previous forms of scholarly communication that it is not easily replicable as a template for continued scholarly communication. In the creation of ‘Emerging Configurations of Knowledge Expression,’ the authors aimed to create an artifact somewhere between the functionality of the traditional written report and the extravagance of Blue Velvet. Accordingly, the digital report consists of seven web pages accessible via http://econfigs.ics.uci.edu. Five topic-based pages complement an introduction and acknowledgement page: “enculturating emergence;” “engaging modalities;” “integrating research spaces;” “interoperaing temporalities;” and “rewarding design.” These topics were identified via content analysis of qualitative data (e.g., semi-structured interviews, conceptual illustrations created by a graphic facilitator, PowerPoint slides, etc.) gathered at the workshop.

Each topic page consists of textual, aural, graphical, and video information. Essays providing an introduction and overview of each topic augment multi-media mosaics comprised of 16 assets. The first mosaic layer includes video interviews conducted at the workshop, slides from workshop participants’ presentations, and graphical facilitator images created during workshop charrettes. At the second layer, each of these assets is accompanied by a textual, aural, or video annotation created by the authors of the report. Finally, through the presence of a ‘submit annotation’ function, users are encouraged to create and submit their own annotations for assets they find interesting. To ensure specific citability, all assets are accompanied by microcitations.

The digital report was opened to the public in December 2012. Following six months of activity, the authors will revise the report to incorporate material included in user-generated annotations. In this way, the authors are exploring a new mode of peer-review. Following the completion of this process, the site will be frozen and archived in the California Digital Library.

Narrative and Engagement in an Interactive, Multi-Modal Digital Report

With the emergence of new communicative technologies come new forms of narrative. Liu (2012) discusses the disaggregation of the book into an entity more closely resembling a playlist. In addition to aspects of Blue Velvet (discussed above), this concept of disaggregation guided the digital report’s structure. By presenting a loose aggregation of five distinct but related topic areas, the report presents a structure that is organizationally functional, but not entirely prescriptive. Users and readers can choose from two paths to follow: 1) the authors’ defined path from one topic to the next; 2) an idiosyncratic, improvised path weaving to and from essays, mosaic assets, and annotations on each of the pages. In either scenario, users can document the details of their narrative journeys through the submission of user-generated annotations. In this way, the report returns the agency of meaning-making and narrative construction to readers/users (Eco, 1979).

Alternative approaches to narrative construction are risky. Readers may not know how to approach a new narrative format, thus decreasing the format’s communicative effectiveness. Developing scholars are often encouraged to read in a specialized way (see Edwards, 2008 for an example). But because of the structure of the digital report, a deeply detailed reading of (at least) some of the assets is requisite – there is no clearly defined introduction or conclusion to skim in order to glean a general
understanding. With greater narrative agency comes a greater obligation to invest time in acquainting oneself with an artifact. It is currently too early to tell what effect the digital report’s structure has on reader engagement; however, careful analysis of use data will be undertaken in the summer of 2013 following the ‘freezing’ of the site.

Conclusion

In early September 2012, the report was disseminated to those who participated in the workshop so that they might have the first opportunities to create annotations and amend the report. A wider public launch was executed in December 2012.

Along with a loose constellation of other scholars and artists (see Goldberg & Hristova, 2007), we have seen the breach of a new, hopeful means of scholarly communication, and we are captivated. Following completion of the initial launch to workshop participants and, subsequently, to the scholarly community at-large, the authors of this poster will watch closely as the interactive dynamics of the report develop.

References

Network Maps and Congressional Frames: Analyzing Bill Titles as a Field of Conflict

Ryan Biava
Department of Political Science
Holtz Center for Science and Technology Studies
University of Wisconsin-Madison
biava@wisc.edu

Abstract
What can we learn about how Members of Congress (MCs) frame the regulation and uses of the Internet by analyzing how they entitle legislation they sponsor that contains the word 'Internet'? Existing literatures on framing argue that MCs, as policy entrepreneurs, utilize fields of public and official discourse available to them to manage conflict within the Congress. Current empirical research of MCs' public communications focuses on long-established and new mediums of communication. This project, however, is interested in a medium of communication available to MCs, but which has received scant attention: the short-titles of legislation they sponsor. Using network analysis of a database of all legislation containing the word 'Internet' from 1994 to 2009 (N=1,170), this project finds that certain terms co-occur more frequently than others, and that the overall structure of co-occurrence demonstrates a coherent deployment of language by MCs along two dimensions: the protection and administration of society.

Keywords: Internet, legislation, network analysis, language, Congress

Extended Abstract
This project seeks to answer the following question: What can we learn about how Members of Congress (MCs) frame the regulation and uses of the Internet by analyzing how they entitle legislation they sponsor that contains the word 'Internet'? Existing literatures on framing argue that MCs, as policy entrepreneurs, utilize fields of public and official discourse (i.e., press releases, policy papers, floor statements) available to them to manage conflict within the Congress and, ultimately, to help pass or defeat legislation. This literature also argues that such framing activity is effective, and can be used strategically to maintain or to change the congressional (or broader political) conflict environment (Baumgartner & Jones 1993; De Boef et al. 2005). In other words, there is a strong incentive on the part of MCs to choose their words wisely.

Some of the earlier research on MCs’ official oratory dates to the mid-20th century, when scholars looked to understand the dynamics of congressional floor debates: who speaks, about which topics, and why (Lehnen 1967). The main thrust of some of this research line focused, later, on applying psycholinguistic research to the analysis of a random selection of floor debate topics in the U.S. Senate, in an attempt to uncover the psychological and ideological orientations of senators (Tetlock 1981).

More current empirical analyses of MCs’ public communications are divided into two main types: those that focus on mediated methods of communication (e.g., television, radio, newspapers) and those that focus on directed methods (e.g., postal mail, personal appearances, web sites; Lipinski 2004). Analyses of the latter have become increasingly easy to perform, given technological evolution; much current research utilizes databases (e.g., the Congressional Record), social-networking services (e.g., Twitter; Mergel 2012), or other online sources (see Golbeck et al. 2010).

The instant research project, however, is interested in a directed method of communication...
available to MCs, but which has received scant attention: the short-titles of legislation they sponsor. The content of this field, like the floor speeches that figure in previous research, is entirely at the discretion of the MC, and the pithy and often clever titles employed demonstrate a high level of thought. For example, following the terrorist attacks of September 11, 2001, the Congress passed and the President signed the USA PATRIOT Act\textsuperscript{1}, of which the short-title was: “Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2011.” Those who work in, or cover the Congress or the state legislatures, will recognize this behavior as commonplace.

This project also advocates a new and as-yet exploratory methodological approach to gain purchase not on how individual MCs are influencing how the public and their colleagues with the words they choose to entitle bills pertaining in some way to the Internet, but instead on the overall structure of terms employed in each Congress (i.e., the two-year term that serves as a temporal and legislative boundary between different congressional sessions). This is done by employing network-analysis methodology, which focuses not on “attributes of autonomous individual units… [but instead] on properties of the social (economic or political) structural environment, and how these structural properties influence observed characteristics” (Wasserman & Faust 1994).

Network analysis has been used in many settings, with the “Deep South” study being an early exemplar (Davis et al. 1941). That study looked at how individual membership in social circles constituted an overall class-caste system. It did so – like all two-mode network analysis – by focusing on individual units (here, people) and the groups to which they belong (here, social circles). Resulting network maps illustrate very well the connections between the units, in the context of the underlying social structures.

The research presented here focuses on individuals and groups, but with a slightly different focus. It looks at the words in the short-titles (i.e., individuals) and the Congresses (i.e., groups) in which they occur. For the period from 1994 to 2009\textsuperscript{2}, there were a total of 1,170 pieces of legislation the full-text of which contained the term ‘Internet.’\textsuperscript{3} This produced a database from which one can ascertain the most common words for each Congress. The resulting network map illustrates the relationship between those words and the underlying legislative structures.

There are several benefits to this research methodology. First, as mentioned above, we can focus on the relationships between terms: do certain terms tend to co-occur more frequently than others? Second, this co-occurrence allows us to ask key political scientific questions: if terms do tend to co-occur, to which political factors should we turn to explain it? Third, two-mode analysis allows us to look at how the Congresses help structure the overall discursive structure, or dominant vocabulary, that comes to define the debate over the Internet.

Prior to reviewing two key findings, it is worth noting that there has not yet been a network map that allows those who are interested in this research field to view the underlying structures described above. Since the research methodology is largely exploratory, its novelty will hopefully give rise to questions which have not yet been asked, and link research fields which have not yet been linked.

Beyond this, there are two key findings, which the poster will highlight and support with network maps and analysis.

First, there are instances where the dominant vocabulary is the same across Congresses controlled by different parties (Figure 1). This provides early evidence that the Internet might not be a topic that conforms to widely supported political science literature on the polarization of conflict within the Congress (Layman & Carsey 2002). Further research could focus on whether the underlying discursive structures are impacted by the party and/or ideological group to which the MC belongs.

\textsuperscript{1} 115 Stat. 272 (2001).
\textsuperscript{2} This time period comprises the 104th to the 110th Congresses. The start-date was determined based on the fact that the Internet was opened, by congressional action, to commercial traffic in 1993. The 104th Congress was therefore the first Congress to be able to treat the Internet as a resource for whatever means its members saw fit.
\textsuperscript{3} Although the network analysis was conducted only on words in the short-titles, the database was constructed by searching the entire body (i.e., the full text) of legislation during the time period concerned.
Second, using statistical information resulting from network analysis, there is a linguistic structure that becomes apparent. Short-titles of bills analyzed tend to break along two key discursive axes: (1) MCs focus on how to protect society, which provides empirical support for existing literature on how early responses to new technologies are often based on fear of that technology; and (2) MCs also focus on using the Internet to reform or administer government and society (Figure 2). This discursive tension could give rise to further research on how political culture can help shape the responses of a society and/or polity to a new technology.

This research, while somewhat exploratory in nature, looks to break some new substantive and methodological ground. It expands upon and brings together diverse research lines, such as congressional behavior and communication; partisan and ideological structures; discourse analysis; and social studies of technology. As such, it is offered up as an opportunity to provoke interdisciplinary discussion and promote inquiry into a key field of official communication that has not often been targeted as such.
Figure 2. Singular value decomposition analysis of the most-frequently occurring terms, showing two factors which explain 64.7% of variance.

References


**Real Talk:**
A Toolkit for Community Engagement, Transparency, and Mobile Governance

**Hale Thompson**  
University of Illinois, Chicago  
lthomp10@uic.edu

**Jes A. Koepfler**  
University of Maryland, College Park  
koepfler@umd.edu

**Katherine Sydenham**  
University of Michigan  
kathejan@umich.edu

**Anthony Hoffman**  
University of Wisconsin, Milwaukee  
hoffma89@uwm.edu

**Abstract**

The *Real Talk* toolkit is a design concept for community engagement, transparency, and mobile governance. It is inspired by the design imperative to close the gap between the way communities are experienced online and in the real world. The toolkit design, which includes vibrantly colored, biodegradable stickers that can be applied to a physical environment, allows a community that is mobilized around an issue to make its concerns, local knowledge, or community pride visible both in the real world and online. Making this information visible enhances opportunities for engagement with local government and community stakeholders.

*Keywords:* e-governance, transparency, mobile phones, ICTs, community informatics

**Introduction**

Declining urban tax bases and the ubiquity of cell phones and other forms of information and communication technologies (ICTs) in the United States have contributed to increasing calls for the development of mobile and e-government platforms (see, for example, [http://codeforamerica.org](http://codeforamerica.org)). A major motivation for such calls is the hope that web-based platforms will lead to increased information sharing and transparency between governments and their constituents. Governments see the potential to collect data about, communicate with, and deliver information to citizens. NGOs and other community organizations see the potential to use such platforms to increase the visibility of their needs, play a larger role in decision-making, and more easily access information that the government collects.

To achieve these goals, mobile-phone-accessible, e-government platforms hold particular promise. Cell phone use has become a ubiquitous mode of communication in the United States. By 2010, 96% of the U.S. adult population owned a cell phone ([U.S. Wireless Quick Facts, 2011](http://www.ctia.org/research-and-data/wireless-industry-data/)). Among young adults in underserved populations, uptake of cell phones is even greater than the general population. Non-white users, particularly Latinos and African Americans, access many more cell phone data functions than their White counterparts ([Smith, 2010](http://www.smithmarketresearch.com)). Nonetheless, a gap remains in terms of continuous access to the Internet and more interactive, user-generated, Web 2.0 functions that may not be accessible via non-smart cellular phones. As increasing engagement in governance by all citizens becomes a greater national priority, it becomes a *design* priority to:

Acknowledgements: National Science Foundation, grant number, etc. and thank you to Amanda Geppert, Dr. Jeffrey Bowker, Dr. Judith Gregory, Dr. Cory Knobel, and the Values in Design 2012 workshop.
Copyright is held by the author/owner(s).
consider technologies that afford access to a range of citizens
recognize varying levels of access to ICTs, and
courage participation in governance from the community up rather than the top down.

Taking these priorities into account ensures that we do not simply bridge divides with technology, but that we build “two-way bridges” of communication and empowerment for communities and local governance (Eglash, 2002). These design priorities were the motivation behind the conceptual design of Real Talk, a multimedia community toolkit for local mobile governance.

In this manuscript, we first describe the methods and design context for the conceptual design of Real Talk, and then detail the key components of the platform. We focus the discussion of this paper on the theoretical and practical implications of the design choices we made, and highlight the potential that Real Talk has for empowering and engaging a broad citizenry through both digital technologies (i.e. mobile phones) and analog technologies (i.e. stickers).

Methods

In August 2012, this paper’s authors formed an interdisciplinary design team of doctoral students from information science, philosophy, public health, and design departments at five different universities as part of the Values in Design (VID) 2012 doctoral workshop (http://vid.ics.uci.edu/2012/02/29/278/). This workshop was an intensive week dedicated to the theory, methods, and design of values in, of, and from information systems and technology.

As part of the “design” component of the workshop, student teams were formed and tasked with a design challenge that explicated both a value and an information technology. Our team used a series of design techniques including defamiliarization (Bell, Blythe, & Sengers, 2005), personas (Kuniavsky, 2003), low-fidelity paper prototyping (Sefelin, Tscheligi, & Giller, 2003), and conceptual investigations (Friedman & Freier, 2005) to engage with the value of transparency and the affordances of mobile devices in response to the design challenge outlined below:

…Transparency in government is linked to better engagement, feelings of empowerment, and participation in local government by the community... Low-income teenagers are more likely to access the Internet through a Smartphone than through a computer... However, these same groups may be among the most underserved in terms of local government access and participation, particularly in the area of laws and services created to protect their rights. Your challenge is to design a mobile application or device to increase transparency of or access to government services to underserved individuals or communities. (Values in Design email, 07/03/2012)

Following Turilli and Floridi (Turilli & Floridi, 2009), our team approached transparency as an instrumental value that enables other values (such as welfare, safety, or accountability) through the making visible of relevant information.

Preliminary Design

The design concept includes a multimedia toolkit that would be funded through public-private partnerships. It is designed to meet communities or neighborhoods where they are in terms of ICT access and use as well as engagement with local governance. The Real Talk concept leverages the current popularity of ICTs and introduces an older form of expression—stickers, although in this case, biodegradable ones—to afford community identification of, conversations about, and resolution of local issues. The stickers also make these issues visible in both the urban landscape and, with the help of ICTs, in cyberspace.

Real Talk reduces barriers to participation by offering incentives. The program provides cell phone minutes or gift cards in order to recruit and retain a core group of community ambassadors (e.g. young adults at a community center). The toolkit provides the community ambassadors with both digital and tangible templates of timelines, budgets, sticker and branding designs, as well as an easy-to-use manual concerning community building, prototyping, and social media tools. Using the toolkit, the
ambassadors deploy five modules: 1) brand development, 2) relationship building and community organizing, 3) a 24-hour “stickout” campaign (see below, Figure 1), 4) documentation of the campaign onto the online community dashboard, and 5) next steps dialogues.

For example, imagine that 10 youth ambassadors in Detroit have designed plain, circular (4” diameter), bright pink stickers (see Figure 1). Using the toolkit for guidance, they decided on this simple brand for three reasons. First, the hot pink color is one that attracts attention. It is uncommon in urban landscapes, and their research revealed that it elicits feelings—calmness and relaxation—that may be conducive to peaceful dialogue. The shape, size, and blank space were found to be aesthetically pleasing with enough room to write messages on if desired. Finally, the simplicity of the design was cost- and time-effective.

During the community-organizing phase of the project, youth discussed the imminent stickout campaign with residents, business owners, organizational leaders, government officials, and law enforcement. These discussions established buy-in among stakeholders who agreed that the campaign may help resolve current concerns around community life. Figure 1 reveals that community members felt that the liquor store’s corner was a source of concern; for example, drug dealing may have been clustering there throughout the day, making its use difficult for passersby and for youth and adults who wait for the bus there. The stickout campaign’s physical and cyber presence (photos and the sticker messages were uploaded to the community dashboard) marked the beginning of the community’s attempt to reclaim urban space, begin dialoguing, and plan for change.

Conclusion

Although other sticker campaigns have been launched to highlight social issues (Chang, 2010; “Project Sticker Shock,” 2003; “THIS IS PUBLIC HEALTH,” 2012), they have not involved multi-modal (e.g. via online, via cell phone, and in community spaces) and multi-directional dialogues (e.g. with each other and with various stakeholders within and beyond the community or neighborhood). They have also not engaged mechanisms like a community ambassadorship and a community dashboard for tracking, locating, compiling, and visualizing the information that the stickers generate. In addition to generating buy-in, the community-organizing phase entails that ambassadors working with community members and stakeholders discuss the possibilities of both positive and negative uses of the stickers during the 24-hour
campaign day and ways to create productive conversations and actions around unexpected outcomes. Following the campaign, stakeholders and citizens assess the information made visible by the stickers and their documentation and determine next steps.

The sticker offers many appealing affordances to facilitate civic participation. Enhancing creative and visual expression, the sticker is considered a more tribal form of communication that began appearing in the 1920s (Vigso, 2010). In the late 20th and early 21st centuries, it has increasingly carried political, and often anti-establishment, messages (Vigso, 2010). The sticker also provides a degree of anonymity and privacy that a purely ICT-driven platform would not, as the only community members gathering and uploading content for the electronic community dashboard would be the community ambassadors. Further, biodegradable stickers are impermanent. While one issue might be relevant for a particular deployment of the Real Talk program, in our design concept, the stickers fade and disappear from the landscape over time, allowing community conversations to evolve as well.

Real Talk is an innovative and alternative community-building model that attempts to bridge the divisions that we often impose upon the public and private spheres as well as online and offline spaces (Gray, 2009). It drives bottom-up social investment, technological capacity building, and community engagement that may lead to participatory indicators of neighborhood-level health, economic, environmental, and infrastructural inequities. Finally, it creates a range of opportunities for low-cost participation in governance and potential payoffs in terms of social equity and engagement.

References


Information Occupation:
Using Information Science to Explore Occupy Wall Street

Oliver L. Haimson
Carnegie Mellon University
oliverhaimson@gmail.com

Josh Cartagena
Rutgers University
jpcartagena@gmail.com

Abstract
Global social movements like the Arab Spring and Occupy Wall Street were unprecedented in their use of social media as an organizing tool and communication technique. To further explore the roles of mobile technologies in the Occupy Wall Street movement, we present an analysis of information diffusion on Twitter as well as describe the relationship between mobility, technology, and spatial practices over a nine-month period. We perform a network visualization and analysis of Twitter data sets as well as utilize ethnographic methods to collect data across several spaces. This paper describes Occupy Wall Street’s practices and describes a model to analyze current and future social movements.

Keywords: network visualization, social computing, social informatics, social movements, technological mediation

Introduction
Social movements during the past decade have taken advantage of ubiquitous computing and social media, facilitating mobility and the flexibility to organize actions and create and share digital content. The same technologies provide researchers with rich insight into these worlds. Understanding Occupy Wall Street, then, requires analyzing protestor practices in a variety of spaces by employing different methods. Citizen journalists, by means of social, mobile media, had a critical role in the Arab Spring uprisings (Tufekci & Wilson, 2012; Wilson & Dunn, 2011), and we find similar implications for OWS. We examine online OWS practices using network and time-series visualizations and analysis of Twitter data, then examine the on-site practices of OWS protestors. Ultimately, we argue that technological mediation (Brewer & Dourish, 2008) gives rise to, and provides researchers new insight into, social movement practices.

Twitter Data Collection Methods

We collected tweets from a list of hypothesized influential tweeters (HITs), including 17 activists/citizen journalists and 5 journalists selected based on their Twitter behavior and OWS involvement. Our HITs’ usernames were used as search terms, allowing us to gather HIT tweets and retweets, as well as retweets and mentions of HITs authored by other users. We also archived tweets using designated hashtags for each of two actions. Data was collected periodically throughout the day on April 1 (March to Commemorate the Brooklyn Bridge Action) and May 1 2012 (General Strike). Our April 1 data set includes 1,077 unique users and 1,609 retweets, and our May 1 data set includes 14,372 unique users.

Acknowledgements: This project was created as part of the i3 Inclusion Institute, University of Pittsburgh. Special thanks to our inspiring i3 leadership and advisory team: Kip Currier, Ph.D., Michael Depew, Courtney Loder, and Joe Sanchez, Ph.D. Thanks also to our teammates Mavis Gyamera and Guadalupe Soto, and to Colleen Jankovic, Ph.D., for proofreading and editing help.


Copyright is held by the authors.
and 42,919 retweets. After collecting the tweets, we scraped the text files and created a from/to matrix based on retweets.

**Network Visualization and Analysis**

Our network visualization and analysis suggests that a Twitter user can be influential and central in a network if she is (1) an information router and/or (2) an information source, and that the two are not mutually exclusive.\(^4\)

A comparatively large number of arrows point toward some HITs (Figure 1), designating them information routers, and away from others, making them information sources. Our results support Lotan et al.'s findings from the Egypt and Tunisian uprisings that journalists are more likely to be information sources, while activists are information sources and information routers (Lotan et al., 2011). We were surprised to find that certain tweeters (e.g., @YourAnonNews) were central information sources in our network, though less clearly related to OWS on the ground or on Twitter. Our network analysis gives insight into what it means to be an influential tweeter, and how information about a social movement can most effectively be spread – by garnering the support of information sources, whose tweets can reach a wide audience, and by employing the help of information routers to further disseminate information.

\[\text{Figure 1. A portion of our directed network visualization of the April 1 action.}\]

**On the Streets and in the Tweets**

Combining Twitter volume with observations from on-site research, we find that Twitter activity occurs most often just before and after an action, and subsides during the most active times of on-site protest. We display the number of tweets during the April 1 and May 1 actions alongside the number of people present at each protest (Figure 2). Peaks in Twitter activity occurred both as the action gained momentum, and as the action drew to a close, suggesting that activists use social media to spread information about social movements during these times. Because relatively little Twitter activity occurred during the time the action was most populated, an effective protest strategy may be to employ a team whose primary role is to spread information via social media, using strategized live-tweeting for instance, during the most active times of an action.

\(^4\) The terms “information router” and “information source” come from Lotan et al., 2011.
Figure 2. Amount of people at OWS actions as compared with related tweets.

**Occupy Wall Street’s Information Ecosystem**

New technologies, technological practices, and forms of mobility give everyday spaces structure and meaning (Brewer & Dourish, 2008). As OWS’s networks flourished, new demonstration locations were selected, and confrontations with the NYPD occurred, protestors adjusted their practices, helping preserve existing meanings while also giving rise to new ones. We briefly discuss the semiotics of OWS, then describe the relationship between protestor practices and the role of mobile devices in different spaces and under separate, ecological conditions.

**Technological Mediation and Social Practices**

OWS’s nonverbal communication developed because a large number of individuals were confronted with, and found ways to negotiate, the same institutional hurdles (Swidler, 1995). N.Y. state law prohibits the use of megaphones without permits, so communication within Zuccotti Park occurred within the constraints of existing encoded social structures (Brewer & Dourish, 2008). In response, nonverbal gestures were created (Figure 3).

**Occupy Wall Street’s Nonverbal Communication**

- Uptwinkles: An upward hand gesture signifying agreement.
- Downtwinkles: A downward hand gesture signifying disagreement.
- Block: Arms crossed in front of chest or in air signifying severe moral/ethical reservations.
- Clarify: A c-shaped hand gesture signifying additional, clarifying information is needed.
- Point of Process: Triangular hand gesture signifying consensus reaching process is not being followed.

---

5 Zuccotti Park is located in lower Manhattan, New York City where OWS protestors camped and lived from September 17 until November 15, 2011. It was renamed Liberty Square during its occupation.
Initially used during Zuccotti Park’s nightly General Assemblies\(^6\), facilitators encouraged their use. As the movement grew in numbers and new spaces emerged\(^7\), these gestures manifested themselves independent of ‘physicality’ or ‘virtuality.’ Conversational quips in local cafés and digital content on the NYCGA website were up twinkled\(^8\) by protestors. What once seemed reserved as a function of decision-making was now used to communicate in different ways and strengthened interpersonal relationships and a sense of shared space, practice, resources, and identity (Baym, 2010).

**Mobile Devices, Varied Roles**

Changes in spatial locations, whether on the Brooklyn Bridge or the NYCGA website, did not alter the signs and social layers (Benyon-Davis, 2007) shared by OWS protestors. Rather, the demonstration’s patterns of movement and action in space were shaped, and shaped by, their cultural logics (Brewer & Dourish, 2008). We describe how technological mediation allowed OWS to encounter and appropriate Zuccotti Park in different ways and how new mobile practices transformed this space as a site of action (Brewer & Dourish, 2008).

Zuccotti Park does not provide Wi-Fi access, however, the Free Network Foundation (FNF) provided protestors with Freedom Towers\(^9\) that created a decentralized, mesh network. Similarly, OWS supporters enabled Wi-Fi tethering on their devices, providing additional Internet access. Protestors and citizen journalists uploaded photographs or livestreamed videos to social media sites capturing group discussions, dances, and individual perspectives. We categorize these mobile and spatial practices under ‘everyday conditions’\(^10\).

Under conditions of duress, these practices change significantly. Network support and information dissemination continue, but rapid organization strategies emerge. This practice was observed during the NYPD raid on Zuccotti Park as tweets and texts were used to contact reporters and organize locations to meet. While helicopter spotlights tracked protestors and NYPD cruisers blocked city streets, protestors used Google Maps to navigate around lower Manhattan. Several hours later, hundreds of protestors and reporters assembled at Foley Square\(^11\) to regroup and document the aftermath.

**Discussion**

Our approach illustrates how network visualization and ethnographic methods are useful when examining how information spreads and describing mobile technologies and the roles they play. As new social movements materialize, understanding the technologies and practices used to appropriate spaces and dominant media narratives will be important areas for future study. While our analysis concentrates on network analysis and technological mediation, similar lenses and methods can be used to perform content and frame analysis (Kitzinger, 2007), or to examine issues surrounding government surveillance of activist communities (Morozov, 2011; Soghoian, 2012).

**References**


---

\(^6\) General Assemblies were nightly meetings where decisions were made ranging from fund allocations to revising proposals.

\(^7\) These spaces include meeting and demonstration locations, Twitter, Facebook, the New York City General Assembly (NYCGA) social networking website, and Tumblr.

\(^8\) In addition to up twinkles (see Figure 4 for description) observed in conversation at local cafés, this gesture was designed and integrated within the NYCGA's website design. It is the equivalent of a Facebook “like” and has a similar upvote/downvote rating function as Reddit.

\(^9\) Freedom Towers are comprised of modems, routers, netbooks, and custom firmware and connected to Clear’s 4G network.

\(^10\) While this abstract highlights two Occupy Wall Street information ecosystem conditions, our poster describes four: everyday, duress, marches, and raid. Under each, we illustrate the devices present and how they are used.

\(^11\) Foley Square is a park located a few city blocks from Zuccotti Park.


A Practice Perspective on Websites for the Sharing Economy

Gabriel H. Mugar
Syracuse University, School of Information Studies
gmugar@syr.edu

Abstract

The sharing economy describes an economic model in which people sell, share, or barter their skills or owned assets directly with others. This economy is facilitated primarily by websites that act as hubs for the visibility and transactions of local assets. Yochai Benkler notes that such peer-to-peer transactions are mediated not by market prices or organizational hierarchies, but by normative frameworks. The ways in which the normative frameworks are produced and perpetuated by transacting parties on the websites has yet to be studied by scholars. This paper proposes a practice perspective as a theoretical framework and Sense-Making as a methodology to explore how users interact with each other on the websites, so as to produce and sustain the normative frameworks critical to the success of the sharing economy.

Keywords: socio-technical, sense-making, collaboration, peer-to-peer, practice perspective

Introduction

The past few years have witnessed an increasing number of websites that facilitate local peer-to-peer transactions of specific assets. Websites like airbnb.com allow homeowners to rent out rooms or entire apartments, zimrides.com allows car owners taking long distance trips to sell seats in their car, and ourgoods.org facilitates connections for artists looking to barter with other artists in New York City. Such websites are part of what Botsman and Rogers call the Sharing Economy (Botsman & Rogers, 2010); peer-to-peer networks where people trade, sell, or share their skills or assets directly with one another. In the sharing economy, the websites act as the meeting place for people with needs to identify people with assets. In addition to matching needs with assets, the websites act as grounds for negotiating the terms of a transaction. As Yochai Benkler points out, peer-to-peer transactions are mediated not by market mechanisms or hierarchical models, but instead by social frameworks (Benkler, 2004). In other words, peer-to-peer transactions are disintermediated, meaning there is no middleman structuring and perpetuating the norms of the transaction. Rather, a social framework of norms upheld by the participants guides the expectations that transacting principals have of each other. As Benkler points out "peoples expectations about the type of interaction in which they are participating can have a huge impact on levels of cooperation" (Benkler, 2011).

The question of how such websites support social frameworks has been addressed in research from a distinctly quantitative perspective, looking primarily at the impact of reputation management systems on the willingness of participants to engage in peer-to-peer transactions on such platforms as eBay (Resnick & Zeckhauser, 2002; Resnick, Zeckhauser, Swanson, & Lockwood, 2006) and Couchsurfing.org (Adamic, Lauterbach, Teng, & Ackerman, 2011; Lauterbach, Truong, Shah, & Adamic, 2009). While reputation management systems are a critical feature of the sharing economy, reputation is only one aspect of what factors into establishing the normative framework of a website for the sharing economy. Other questions that are important to understand for the continued success of the sharing economy include, for example, how users engage features of the site to help them understand how to effectively negotiate the price or trade of assets, how the site's features help users to communicate effectively with other participants on the site, or what users look for when assessing the reputation of other users.

Acknowledgements: The author would like to thank Nilan, M., Tonkinwise, C., and Benkler, Y. for their feedback.


Copyright is held by the author/owner(s).
This paper proposes expanding the scope beyond reputation by using the practice perspective in social theory to provide a richer description of how users negotiate and establish norms of transactions with other users. The proposed framework also seeks to contribute to existing literature on online collaborative environments by not only introducing the sharing economy phenomenon to the literature, but by also exploring social norms in collaborative environments from a perspective of how users enact and sustain normative frameworks through their interactions with human and non-human actors.

The Practice Perspective

The practice perspective in contemporary social theory views social reality as being comprised of a multitude of practices, or specific activities that, through repetition across time and space, represent pockets of social order, or institutions. Each practice is thus defined as a framework of social norms that all individuals who partake in the practice share. Thus, in the view of the practice perspective, collaboration or interaction of any kind between social actors is contingent on a shared frame of reference regarding the social norms of that practice.

Frames of reference, or what Bourdieu describes as Habitus, are "the universalizing mediation which causes individual agent's practices...to be nonetheless "sensible" and "reasonable" (Bourdieu, 1977). If social actors do not share a common frame of reference or habitus, their actions will not be intelligible to one another. Shared frames of reference are also where the idea of social order exist for practice theory. Social order then is understood as shared expectations about how actors will behave in a particular social setting.

The perpetuation and co-construction of common frames of reference are tied to what Bourdieu describes as embodied or objectified modes of practice production. In the case of embodied modes of practice production, people refer to their memories of similar situations to inform present action, or they may mimic the actions and appearances of those around them in that moment. In the case of objectified modes of practice production, habitus is not reliant on interpersonal interactions in order to be perpetuated; rather, symbolic objects represent the key aspects of a habitus. Bourdieu uses the example of calendars as objects used to perpetuate particular agricultural activity, where the calendar as object works as a cue to inform the organization of agricultural work (Bourdieu, 1977). In either mode of practice production, the perpetuation of social order is a function of what different social theorists view as an inclination of social actors to always reflect on their social settings so as to establish common ground with those around them (Giddens, 1984; Goffman, 1974).

This idea of being reflective in the context of both embodied and objectified modes of practice production should evoke questions about the user experience in websites for the sharing economy. For example, what social cues do users look for in each other? What role does the design of the website play in helping to support a common frame of reference between transacting parties? Attention to such questions will help provide a richer perspective of how the social framework of websites for the sharing economy are negotiated and perpetuated across time and space.

Research Method

This paper proposes the use of Dervin's Sense-Making as a methodology to unpack the activity of users of websites for the sharing economy in order to understand how they produce and establish a common framework for collaborating. Sense-Making comes out of research on information behavior, looking at the information seeking practices of information systems users. In this case "Information seeking and use are seen as processes, the step-taking that human beings undertake to construct sense out of their worlds" (Dervin, 1992).

Sense-Making is grounded in the ontological assumption that movement through life involves persistent confrontations with discontinuity, meaning that social actors constantly find themselves in moments where they need to make sense of the current situation in order to proceed (Savolainen, 1993). Sense-Making then is a description of the activity of "reflexive monitoring of the environment of action"(Savolainen, 1993). Like the practice perspective in social theory, Sense-Making views social order as the product of an ongoing process perpetuated by social actors as they work together on "the development of shared rules and procedures for meeting fundamental life discontinuities" (Savolainen, 1993).
One of the more prominent methods for Sense-Making is the Timeline technique in which the interviewer asks the respondent to describe a problem situation step-by-step and then at each step, describe any challenges they may have experienced and what either helped the respondent move past the challenge or what held them back (Savolainen, 1993). In the case of websites for the sharing economy, because the perpetuation of normative frameworks are critical to successful transactions, the Timeline technique would allow for a detailed description of the ways in which users interact with each other and with the features of the site in order to establish the common ground needed for successful transactions. Such a description would help researchers interested in the ways in which normative frameworks are perpetuated in online collaborative environments.

Conclusion

The sharing economy is reliant on normative frameworks that establish the common ground social actors require when transacting with one another (Tonkinwise, 2011). Currently the majority of research on the sharing economy looks at cause and effect relationships between features of such sites and participants’ willingness to participate. While this helps establish some of the reasons why people are willing to participate in the sharing economy, it does not address the "microprocesses of social life"(Savolainen, 1993), meaning that current research does not show how the features work to mediate the relationships between transacting parties.

The proposed theoretical framework and research method has the potential to contribute to a number of fields. First, the field of Computer Supported Collaborative Work has scholars who engage the question of how design of online collaborative environments work to inform users about the particular normative structure of the environment (Erickson, Halverson, Kellogg, Laff, & Wolf, 2002; Harrison & Dourish, 1996; Harrison & Tatar, 2007). Such work for the most part has remained at a theoretical level and is in need of opportunities for more empirical investigation. The rise of websites for the sharing economy provides opportunities for such investigation. For designers, the proposed approach may help to uncover not only how users are establishing and perpetuating the social norms for a specific site, it might also show designers what social norms are important to users.

References


Cultural Context of Social Media Use: How Do Iranians Adapt and Use Facebook?

Emad Khazraee
Drexel University
emad@drexel.edu

Abstract

It has been discussed intensively that technology can shape social and in return is being shaped and adapted by its social context of use. Social media technologies have become pervasive in the recent years. It has been discussed in the literature that even very rigid technological artifacts are subject to modifications in different socio-cultural contexts of use. Social Networking Sites (SNS) have become pervasive in the recent years. It is important to understand how technologies such as SNS are being adapted in different cultural contexts. This study concerns adoption, use and adaptation of Facebook among Iranians and reports the preliminary findings of a qualitative study of Facebook use among Iranians. We observed that the use of SNS helps maintaining parallel social realities among Iranians. Although, the use of SNS always involves concerns about security threats from the state.

Keywords: social media, social networking sites, cultural context, technology adaptation, Facebook

Introduction

The interplay between social and technical has been the subject of intensive studies (Bijker & Law, 1992; MacKenzie & Wajcman, 1999). It has been discussed that technology can shape social and in return is being shaped and adapted by its social context of use. Social media technologies have become pervasive in the recent years. It has been discussed in the literature that even very rigid technological artifacts are subject to modifications in different socio-cultural contexts of use (Akrich, 1992).

Social Networking Sites (SNS) have become pervasive in the recent years. Today, more than 840 million users around the world from different socio-cultural backgrounds use Facebook on a universal platform. In some cases, technologies are used differently from the initial purpose of their design (Gordon & Killick, 1993). Thus, it is important to understand how technologies such as SNS are being adapted in different cultural contexts. The study of the adaptation and use of such technologies can reveal overlooked aspects of social media in distant cultural contexts. This study concerns adoption, use and adaptation of Facebook (as one of the most popular social networking sites) among Iranians. This poster reports on the preliminary findings of this study.

Data Collection and Methods

Collecting data about online social practices on Facebook is not easy for two reasons: First, an important aspect of such social practices is considered private and cannot be observed publicly. Second, companies such as Facebook are reluctant to share data. In this study, qualitative interviews were utilized to understand how Iranians use Facebook.

Considering these limitations, this study followed a stratified quota sampling strategy which is a common practice in qualitative research especially when looking at a small number of subjects (Lindlof & Taylor, 2002; Patton, 2002). A recent major study on the General Media Use (GMU) in Iran (Wojcieszak,

---

Copyright is held by the author.
Smith & Enayat, 2012) provided general information about media consumption in Iran. The sample for the present study stratified based on the statistics and demographics derived from GMU data set. The aim was to reflect a similar socio-demographic population of SNS users as identified by the GMU study.

For this study, eight users were interviewed. Interviewees were selected from a list compiled through a snowball sampling strategy. The interviews followed a semi-structured model (one face-to-face and seven phone interviews utilizing Skype). Each interview lasted for about an hour. The interviews were recorded and fully transcribed. The preliminary analysis then has been done using Atlas.ti Qualitative Data Analysis software.

Table 1
Demographics of participants of the study

<table>
<thead>
<tr>
<th>Interviewee</th>
<th>Gender</th>
<th>Age</th>
<th>Education</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>Female</td>
<td>23</td>
<td>Undergraduate degree</td>
<td>Outside Iran</td>
</tr>
<tr>
<td>Two</td>
<td>Female</td>
<td>27</td>
<td>Undergraduate degree</td>
<td>Iran</td>
</tr>
<tr>
<td>Three</td>
<td>Female</td>
<td>32</td>
<td>Undergraduate degree</td>
<td>Iran</td>
</tr>
<tr>
<td>Four</td>
<td>Female</td>
<td>21</td>
<td>High school</td>
<td>Iran</td>
</tr>
<tr>
<td>Five</td>
<td>Male</td>
<td>37</td>
<td>Undergraduate degree</td>
<td>Iran</td>
</tr>
<tr>
<td>Six</td>
<td>Male</td>
<td>29</td>
<td>Graduate degree</td>
<td>Iran</td>
</tr>
<tr>
<td>Seven</td>
<td>Male</td>
<td>23</td>
<td>High school degree</td>
<td>Iran</td>
</tr>
<tr>
<td>Eight</td>
<td>Male</td>
<td>22</td>
<td>High school degree</td>
<td>Iran</td>
</tr>
</tbody>
</table>

Discussion

Given the size of sample and the exploratory nature of this study it is too early to claim about the findings; however the following themes have been observed during the analysis of the data from interviews:

**Parallel realities:** Facebook plays the role of an extension to the social life of many Iranians. This online social life makes it possible to practice those social norms and forms that have been repressed by the official dominant discourse in society. The possibility of maintaining these parallel realities creates very complex social relationships and affects the use of Facebook among users.

**Parallel realities require parallel social representations:** An interesting theme appeared in the interviews is that because of two parallel realities in Iranians’ social life they need multiple social representations. For example, in workplace one has to follow the mandatory rules to wear hijab or specific dress code, but in the familial relationship and more private spaces one do not follow those rules. This subtle issue has been ignored in the Facebook design. You can have just one profile picture at a time based on this assumption that you have one social representation. This can be extended to other Muslim cultures as well. Females in Islamic culture have two level of presentation. One should wear hijab in public but can be more relax in the familial settings.

**Security concerns and spying stories:** All of the participants confirm that they know about Facebook surveillance by Iranian authorities. Seven out of eight interviewees indicated that they knew stories that friends or acquaintances faced threats from the state because of using Facebook.

**Security concerns and disguised identities:** Because of the concerns about the use of Facebook, Iranian users become more conscious regarding adding people to their friend lists, they try to confirm the authenticity of the request via other routes such as phone call or email. The number of mutual friends seems not to be a valid strategy to accept friendship requests anymore. Also, many users to protect themselves experienced the change of their name, use of non-distinguishable profile picture and restrict their account search-ability.

**Concerns about Facebook design and violation of privacy:** While most of users were optimistic about improvements in Facebook privacy settings still there are complains about the violation of privacy by Facebook, for example some users consider the side tracker feature in Facebook as a violation of their privacy by revealing their activity that they are not willing to share with others.
**Facebook can be used for business in Iran:** Some users mentioned business uses of Facebook such as fashion design, wedding ceremony services, wedding make-up services, photography services, and information dissemination about events such as music concerts and public events. Considering Facebook as an illegal venue it is interesting how people use Facebook for these purposes and overcome the risks of use.

**Facebook as a source of tension among generations:** Facebook can create tension among younger and older generation mostly based on the image propagated by the state. Due to GMU report 92% of Facebook users are below 40 years old, thus few people over 40 are on Facebook and have a clear image about it. Especially, religious families are more pessimistic about the use of Facebook.

**Facebook as an important channel to stay connected with expatriates:** One of the main functions of the Facebook for all users is to stay connected especially with friends and family abroad. This platform provides a rich channel of information without having direct contact that you can follow your friends, their photos and achievements. Also, it provides a rich channel for young Iranians to learn about living abroad via looking to the expatriate’s life on social media. However, people mostly share their special moments on Facebook (vacations, parties and travels) which may create a biased unrepresentative image of joyful lives abroad.

**Table 2**  
A Summary of Facebook Use among Interviewees

<table>
<thead>
<tr>
<th>Interviewee</th>
<th>No.1</th>
<th>No.2</th>
<th>No.3</th>
<th>No.4</th>
<th>No.5</th>
<th>No.6</th>
<th>No.7</th>
<th>No.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Spy Story</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Circumvention Tools</td>
<td>Proxy tool</td>
<td>UC Browser</td>
<td>Proxy tool</td>
<td>Proxy tool</td>
<td>VPN</td>
<td>Proxy tool</td>
<td>UC Browser</td>
<td>Proxy tool</td>
</tr>
<tr>
<td>Approx. Time spent daily on Facebook</td>
<td>Always Online</td>
<td>2 hours</td>
<td>2 hours</td>
<td>1.5 -2 hours</td>
<td>30-40 min</td>
<td>30-45 min</td>
<td>Always Online</td>
<td>30 min</td>
</tr>
<tr>
<td>Connection Type</td>
<td>Broadband</td>
<td>ADSL 256 Kbps</td>
<td>ADSL 512 Kbps</td>
<td>ADSL 256 Kbps</td>
<td>ADSL 128 Kbps</td>
<td>ADSL 256 Kbps</td>
<td>3G</td>
<td>ADSL 1 Mbps</td>
</tr>
<tr>
<td>Searchable</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>-</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>Using groups to share content</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Writing language</td>
<td>Pn</td>
<td>En</td>
<td>Fa</td>
<td>Fa</td>
<td>Pn</td>
<td>En</td>
<td>Fa</td>
<td>En</td>
</tr>
<tr>
<td>Mobile access</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

* Interviewee No. 1 is living in the US now
** Proxy tools mentioned in the interviews are: Freegate, Sciphone, Web freer, Ultrasurf, Your-freedom and Cocoon
*** Pn: Penglish (using English alphabet to transliterate Farsi), En: English, Fa: Farsi, Tu: Turkish

**Future Works**

To confirm the findings of this study it is necessary to test the observations by a larger sample. This study provided the basis to design a survey to confirm the findings. The next step to pursue this study will be administering an online survey among Iranian Facebook users to collect quantitative about their use of Facebook.
References


Characterizing the Scholar H-index Via Full-text Citation Analysis

Star X. Zhao  
Department of Information Resource Management,  
Zhejiang University,  
Hangzhou, China  
starzhao@zju.edu.cn

Xiaozhong Liu*  
School of Library and Information Science,  
Indiana University,  
Bloomington, IN.  
liu237@indiana.edu

Fred Y. Ye  
School of Information Management,  
Nanjing University,  
Nanjing, China  
and Department of Information Resource Management,  
Zhejiang University,  
Hangzhou, China.  
yy@zju.edu.cn

Abstract

This study proposes a method to characterize the scholar h-index by full-text citation analysis. The method combines the citation context analysis, graph mining, and supervised topic modeling to modify the oversimplified process of citation count, and provides more sophisticated assumptions for the scholar h-index in two aspects: the context of citation and the supervised topic-related measure.

Keywords: full-text citation analysis, h-index, academic impact, text mining, bibliometrics

Background and Objective

The bibliometrics is an important means to characterize scientific publication, scholar, or domain. In this field, the h-index (Hirsch, 2005), considered as one of the most renowned and successful bibliometrics indicators in recent years (Egghe, 2010; Norris & Oppenheim, 2010), has been applied in many aggregative levels (e.g. scholar/journal/institution/country/ science funding) and extensive fields (e.g. finance/network) (Byström, 2011; Zhao & Ye, 2012). Among them, Scholar h-index is the earliest and most used application. A scholar with an index of h means that he/she has published h papers each of which has been cited at least h times. This measure balances the number of the scholar’s high cited papers and the number of citations. In the sense of citation analysis it reveals both two significant aspects of scholar’s published works: productivity and impact.

However, a major limitation of traditional citation analysis is that the classical method is focusing on citation counts, while ignores the context, topic or motivation of citations. There are different reasons of citing a paper, such as identifying origin, introducing methodology, providing background, giving credit, criticizing others’ work and addressing the interestingness (Garfield, 1964; Liu & Rousseau, 2012). However, most previous h-index implementations ignore most of these qualitative features. Another disadvantage of traditional citation analysis occurs as simplifying the multi-citing to one citing. For instance, if paper A cites two or three different texts of paper B, the citations between A and B will just be simplified to one linkage between them. However, intuitively, based on citation frequency and citation context (or topic), citing paper’s credit should NOT be evenly distributed to the cited publication, and some citations should be more important than others (Voos & Dagaev, 1976; McCain & Turner, 1989; Liu, Zhang & Guo 2012).

For scholar’s h-index, these limitations lead to essential problems of its validity and reliability. First, the scholar’s h-index is constructed by the ambiguous citing meaning and topic, i.e., the indicator oversimplifies the citation relationship because a cited paper can make essential or trivial contribution to the citing paper. Second, the calculation of scholar’s h-index omits the multiple citing between two papers. In most circumstances the multiple citing indicates their close relevance, thus this process seems not fair and might lose some important information.
In this proposed research, based on Liu, Zhang and Guo (2012), we employed a supervised topic modeling algorithm, Labeled LDA (LLDA) (Ramage et al., 2009), to infer the publication and citation topic distribution, where each topic is a probability distribution of words and the label of the topic is an author contributed publication keyword. The publication and citation topic probability distributions, then, can be converted to the vertex (publication) prior and edge (citation) transitioning probability distributions to enhance citation network PageRank (with prior distributions) for calculating topical h-index. More specifically, we assume that words surrounding a target citation (citation context) can provide semantic evidence to infer the topical relevance or reason for the target citation, and that a citation network with prior (topic) knowledge can enhance classical bibliometric analysis, i.e. based on the citation context, if a cited paper contributes to the core topic(s) of the citing paper, this cited paper should get more credit from the citing paper (higher transitioning probability). Because each vertex or edge on the citation network is associated with a topic probability distribution, the enhance PageRank can generate an authority vector, and each score in the vector tells the publication or author topical importance, which will, then, be used to calculate author topical h-index.

Methods and Designs

The method proposed in this study combines the citation analysis and text mining to replace the oversimplified process of citation count. It applies a supervised topic modeling algorithm (Labeled LDA) to produce the publication and citation topic distribution, where each topic label is an author contributed keyword and each topic consists of a probability distribution of words. Then, a weighted citation network can be constructed by the publications (nodes) and citations (edges) according to their topic probability distributions. This method is based on that assumptions that: 1) for a target citation, surrounding words (context) can reveal the citation topical motivation; 2) a cited paper which contributes to the core topic(s) of the citing paper should obtain more weights (credits) from the citing paper; and 3) the publication (node) importance can be scored by the citation network which is associated with a topic probability distribution.

For the new scholar h-index based on the full-text citation analysis, we attempt to use the following steps to measure the author topical importance:

1. In a paper set, analyze all the full text of the paper, and extract all the topics along with their topic labels (author provided keywords).
2. Construct the weighted citation network, \( \mathcal{G} = (V, E) \), with two kinds of prior knowledge: publication topic prior and citation topic transitioning probability distribution.

Each vertex, \( v \in V \), on the citation graph represents a publication, with the publication topic prior probability vector \( \{p_{z_{1}}(v), p_{z_{2}}(v), ..., p_{z_{K}}(v)\} \), where \( p_{z_{k}}(v) \) is the prior probability of vertex \( v \) for topic \( z_{k} \), and \( \sum_{k=1}^{K} p_{z_{k}}(v) = 1 \). Each edge, \( e \in E \), on the graph represents a citation connecting \( v_{i} \) and \( v_{j} \) (\( v_{j} \) cites \( v_{i} \)). The topic transitioning vector for each edge is \( \{p_{z_{k_{1}}}^{x}(v_{i}|v_{j}), p_{z_{k_{2}}}^{x}(v_{i}|v_{j}), ..., p_{z_{k_{K}}}^{x}(v_{i}|v_{j})\} \), where \( p_{z_{k_{j}}}^{x}(v_{i}|v_{j}) \) is the probability of transitioning from vertex \( v_{j} \) to \( v_{i} \) for topic \( z_{k_{j}} \).

3. Compute each scholar’s h-index by employing publication topic distribution \( p_{z_{k_{j}}}^{x}(v) \) and citation transitioning probability \( p_{z_{k_{j}}}^{x}(v_{i}|v_{j}) \).

By this new method, a topic-related scholar h-matrix can be set up, as shown in Fig.1.
There are three merits for this proposed research. First, Labeled LDA, used to characterize publication and citation for this research, is a supervised topic model that constrains LDA by defining a one-to-one correspondence between LDA’s latent topics and user tags (keyword metadata). Labeled LDA can directly learn word–tag correspondences, which has been demonstrated to improve expressiveness over traditional LDA with visualizations of a corpus of tagged web pages. It is a promising method to model topics for h-index, and which could be used to optimize the ranking algorithm, and important for result evaluation and interpretation.

Second, unlike classical scholar h-index, our method produces topic based on author h-index scores. Namely, for different topics, a specific author could have different h-index scores. Consequently, the scholar h-index can be compared in the same research topic. It provides much fairer results for the scholars who involve multiple topics or fields.

Last but not least, this new method, considering full text publication and citation transitioning probabilities, may favor authors that make significant contributions but which have not yet received many citations. For instance, our method will grant more credits to new papers and unknown authors if they are making essential contribution to important (high cited) publications. This is very important for academic information retrieval and recommendation systems also.

Dataset and Evaluation

We used 41,370 publications from 111 journals and 1,442 conference proceedings or workshops on computer science for the experiment (mainly from the ACM digital library), where the full text and citations were extracted from the PDF files. The selected papers were published between 1951 and 2011. From these we extracted 28,013 publications’ text (accounting for 67.7% of all the sampled publications), including titles, abstracts, and full text. For the other publications, we used the title, the abstract, and information from a metadata repository to represent the content of the paper.

In order to evaluate our work, we will sample a list of topics (with keyword labels). Domain expert will sample some main conference proceedings or journals for each candidate topic. By using classical and this innovative h-index method, we will 1) identify the most important authors from this community; and 2) predict the most important authors (not yet important) in a number of years. MAP and NDCG indicators will be used for this evaluation.

Outlook

Our methods attempt to provide more appropriate assumptions for the scholar h-index in two aspects: the context of citation and the topic-related measure. In future works, we will implement the ideas and designs by using ACM data. We believe that the citation measures should consider more details of the context, and the full-text mining would be a potential tool for this purpose. Theoretically, these methods can be applied for h-index at other aggregative levels also, such as journal, institution or...
research field. Although there are still many difficulties to understand and interpret the semantic or motivation of citations accurately and completely, the full-text citation analysis provides the primary insight to observe and characterize the context of citations.

References


Abstract

Using survey tools to develop and implement community outreach plans, librarians and information professionals in rural Texas found out what their constituents wanted and built community-centered outreach programs. Community-centered outreach programs gained more institutional support and increased competitiveness of rural information professionals.

Keywords: continuing education, information profession, rural development, community outreach, community-centered program

Introduction

According to the US Census data, eighty-four percent of land in the United States is rural with twenty percent of the total population living in rural areas (United States Census Bureau, 2010). Government report also suggested that baby boom residents are now poised to significant increase in rural and small-town America, with major social and economic implications for their chosen destinations (Cromartie & Nelson, 2009).

Users in rural and small town America are experiencing various challenges due to demographical changes. Traditionally public libraries have been perceived as a necessary fixture in a vital and vibrant community. Public libraries were once viewed as an indicator of what makes a particular community attractive to existing and potential residents. With the advent of certain technologies in recent years, such as the Internet, smart phones, and e-readers, this sense of libraries being an anchoring element of the community has been eroded. The perception by many governing officials that libraries are no longer necessary since “everything can be found on the Internet” has weakened support and funding for libraries. When fiscal cuts are needed, the books and information resource budget, staffing, and facility maintenance are usually viewed as nonessential expenses exposing such expenditures as easy places to cut.

At the same time, rural libraries and information centers, those serve populations under 25,000, are facing a shortage of information professionals with professional degrees and experience on library and information services. Rural librarians’ professional skills are usually limited both by library practice and the resources available (Vavrek, 1995). Even those with master’s degree may found local community needs do not meet their specific leaning objectives (such as reference or cataloging) at library and information science programs (Williams, Bishop, Bruce, & Irish, 2012). The lack of an educated workforce in rural areas is directly related to the dearth of direct funding and lack of securing community support and commitment to rural libraries.

One way of gaining more community support is for libraries to design and implement community outreach activities that partner with outside organizations. However, community engagement has been ignored in information science curriculum. No iSchools or ALA-accredited library science programs offer courses related to community outreach. To address this important issue, this poster explored the following questions:

Acknowledgement: This project was supported by Robert and Ruby Priddy Charitable Trust.
1) To what extent will community outreach activities enhance the support of rural libraries?
2) To what extent will community outreach activities enhance professional networking of rural libraries?
3) To what extent will community outreach activities enhance professional development of rural information professionals?

Previous Studies

Community Outreach

Traditionally, community outreach refers to equitable library services that are offered outside the walls of the library building such as bookmobiles, homebound services, deposit collections, and service to residential facilities, hospitals, shelters, jails, and other target populations (Meadows, 2004). In recent years, community outreach means an activity that connecting, delivering, partnering, collaborating, teaming up, joining forces, fostering engagement outside of library entities (Barco, 2010). Most importantly, it involves a process of building trust and partnerships with those outside the library (Snell, 2012).

Experiential learning

The essence of community-centered program is experiential learning. Experiential learning theory (ELT) provides a holistic model of the learning process and adult development. Experiential learning theory emphasizes the central role that experience plays in the learning process. The theory originated in the previous works of Karl Jung (1938) and Jean Piaget (1985), and developed by David Kolb (1984). According to David Kolb (1984), learning takes place through four steps: (1) concrete experience followed by (2) observation and experience followed by (3) forming abstract concepts followed by (4) testing in new situations.

While members of the library staff may attempt to interact at some level with local community leaders and organizations, these interactions are often sporadic, random, and without a coherent plan of action. This is representative of the lack of methodology and resources for information professionals in rural settings.

Methods

This study applied One Group Pretest-Posttest Design with several cohorts of librarians or information professionals from rural small towns of Texas. One Group Pretest-Posttest Design is one kinds of pre-experimental design that is applied when true random assignments cannot be achieved (Powell & Connaway, 2004, p. 179). Seventy-four librarians and information professionals participated in this study. Participants first completed a pretest survey on the demographic statistics and current outreach actives. Most of the participants work in a library or information center with an average full time staff of two or less, and serve a community less than 2,500 of population. Among them only five have earned an MLS degree, and most of them earned a bachelor’s degree or some college credit.

The experiment followed four stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. During the first stage, participants conducted surveys to understand their community needs. After consolidating with local statistics from CENSUS data, they actively reflected community needs. This step encouraged them to think user-centered approach in library services. The abstract conceptualization step was the most difficult part. During the period rural librarians designed community outreach plans and then implemented them in step 4. With the help of project handouts and mentorship from graduate students, librarians followed the templates and finished the draft plans. Eventually plans were implemented and reported to the project managers.

After the participants completed the whole learning process, they were invited back to the site for a two day workshop. Participants completed a posttest survey followed by a focus group interview. The interview session was recorded and transcribed to text. The whole process took from 3 to 6 months.
Results and Conclusions

Initial analysis of data revealed the success of community-centered outreach and partnership programs. Pretest and posttest data shows gains on perception of community support to libraries. Community-centered activities helped their professional growth, enhanced the presence of information services in communities, and greatly increased financial and non-tangible support to libraries and information centers.

In conclusion, successful community outreach may help to change the image of libraries and information centers as a necessity rather than a nicety in today’s technology driven society. In the past information professionals often proceeded from the point of view of the workflow, their needs, wants, and priorities. Services were offered with little to no public user input. It seems valuable to train rural information professionals, most of whom without a library science or information science degree, effective methods on how to find out community needs rather than deciding unilaterally what programs or services they want to offer. Community-centered programs have proved to gain broader support and increase the competitiveness of rural information services.

References


Exploring Cyberbullying Through Visual Narratives

Leanne Bowler
University of Pittsburgh
lbowler@sis.pitt.edu

Eleanor Mattern
University of Pittsburgh
lbowler@sis.pitt.edu

Cory Knobel
University of California Irvine
cknobel@uci.edu

Patrick Keilty
University of Toronto
p.keilty@utoronto.ca

Abstract

This study examined the use of visual narratives as a method to explore young people’s views about cyberbullying. Two focus groups—a group of five teens in high school and a group of undergraduates—used storytelling and sketching to frame their perceptions of cyberbullying around a narrative and to propose design features that might afford young people the time to pause and reflect on their actions in social media before they participate in cyberbullying.

Keywords: social media, cyberbullying, adolescents, narrative inquiry, design

Introduction

In the hands of young people who are still developing their impulse control and are particularly vulnerable to peer-pressure, social media can allow for “online expressions of offline behaviors” and facilitate negative and damaging activities, one of which is cyberbullying (O’Keefe et al., 2011, p. 800). This poster abstract reports on a study that examined the use of visual narratives to explore young people’s views on cyberbullying and for generating design interventions that would combat mean and cruel behavior online. Since social media technologies both embed social values, as well as provide a platform for the constraint and enablement of personal value expressions, we saw these two areas as promising indicators toward technological redesign (Friedman 1996, Knobel 2011).

Two focus groups—a group of five teens in high school and a group of undergraduates—used storytelling and sketching to frame their perceptions of cyberbullying around a narrative and to propose design features that might afford young people the time to pause and reflect on their actions in social media before they participate in cyberbullying. Four “cyberbullying stories” were constructed by the participants, each one revealing two sub-plots—the story that “is” (as perceived by these participants) and the story that “could be” (were certain design interventions embedded in social media). Our focus in this poster abstract is on the story that “is”.

Background

Bullying is a historic problem, but 21st century technologies have introduced new tactics for aggressive behavior (Juvonen and Gross, 2003). “Cyberbullying,” as a distinct form of bullying, has consequently entered the vernacular, with scholars characterizing cyberbullying behavior as the intentional use of technology as a means to hurt another individual. While the traditional discourse surrounding bullying has suggested that it is merely a rite of passage for young people (Elinoff et al., 2004) the scholarly literature and popular media reveal that bullying behavior both on- and off-line may have lasting and devastating consequences. Juvonen, Graham, and Schuster (2008) characterize bullying as a “major public health concern facing youth,” describing adjustment difficulties, mental health challenges, and violent behavior as among its effects. The online environment provides an apposite set of

Acknowledgements: We gratefully acknowledge the participation of the young people in this study. Their contributions are invaluable.


Copyright is held by the authors.
factors for bullying to occur. Current research in online behavior and cyberbullying suggests that people with depression (which often affects bullying perpetrators and targets) tend to prefer online social interaction, which may drive more behavior into the “cyber” context of bullying (Caplan 2003).

Efforts have been made to combat cyberbullying through intervening measures. The literature suggests that these interventions can be classified into three types: 1) law and policy; 2) curriculum and campaigns; and 3) technological responses. This study explored the third category of interventions, technological responses, and employed narrative inquiry to prompt user-generated narratives and design solutions.

Methodology

This study examined the use of visual narrative inquiry to explore cyberbullying and design interventions. A qualitative methodology most commonly employed in education research, Connelly and Clandinin (2006) define narrative inquiry as “the study of experience as story” (p. 375). Researchers who employ this approach are united in viewing the narrative as both the object of study and the method (Connelly & Clandinin, 1990). Our implementation of narrative inquiry departed from the general use of this approach as a means to uncover the experience of our participants and instead probed our participants’ perceptions of the cyberbullying experience as they imagined it would be for someone else. This was done so as to avoid potentially embarrassing any participants who might not have wished to self-identify as victims of cyberbullying.

Connelly and Clandinin (1990) explain that the techniques for data collection in narrative inquiry vary and might include field notes, journals kept by participants, and interviews. Bach (2008) speaks to the use of visuals as a data source in narrative inquiry and characterizes this process as “visual narrative inquiry.” While Bach’s focus is on photography as an added layer to a narrative, we used sketching as a data collection technique. Tversky (2011) argues that visual communication is a valuable complement to spoken words, conveying information and meaning through “marks on a page, virtual or actual, and spatial relations, proximity and place on a page” (p. 502).

Procedures

Two focus groups were conducted in the spring of 2012. The first focus group consisted of five female undergraduates: two sophomores, two juniors, and one senior. The second focus group consisted of three boys (ages 14, 15, and 17) and one girl (age 15). Convenience sampling was used to recruit participants.

The first session with the undergraduate students lasted three hours, with equal portions of storyboarding and group storytelling. Realizing that three hours may be too long for teens, we shortened the protocols with the teen group, focusing more on the sketching and less on the group storytelling. In addition to the participants, there were three investigators in the classroom. Two of the investigators interacted with the participants and the third investigator observed and took notes.

The two sessions were divided into five parts:

a) Setting the Stage: Round of introductions and a “setting the stage” discussion.

b) Storyboard/Sketching activity: Participants divided into small groups of two or three and then, together, sketched a storyboard that told a story of cyberbullying. Participants decided to situate their story within Facebook. Four large sheets of self-adhesive paper were stuck side-by-side to the walls of the classroom, creating a canvas for the storyboards.

c) Sticky-note Activity: Participants were asked to draw on sticky notes an object or an action that might make people in their story stop and think. Both groups were asked to add a second note explaining why the intervention might work. Notes were placed on the large sketches at a point in the narrative showing where the intervention would happen.

d) Group Discussion: Participants and investigators gathered around the sketches and each group narrated their story. The investigators asked clarifying questions. The discussion was audio-recorded and later transcribed.

e) Debriefing: Participants were asked for feedback on the study’s methods.
Results

The combination of narrative inquiry, sketching, and group discussion proved to be a powerful, non-threatening tool for revealing aspects of the participants’ cyberbullying stories. Four stories were constructed by the participants, each story revealing two sub-plots – the story that “is” (as perceived by these participants) and the story that “could be” (were certain design interventions embedded in social media). Our focus in this abstract is on the story that “is”. Figure 1 below shows one of the storyboards completed by an undergraduate group.

![Figure 1: A cyberbullying story](image)

Preliminary findings

- The use of narrative inquiry might suggest that the participants would produce linear plots with a clear beginning, middle, and an end. However, the “plots” were messy, littered with dead ends, unresolved problems, and indeterminate conclusions.
- These were stories that featured a wide cast of characters, revealing a social network that went well beyond the binary relationship of “bully/victim”, a simple narrative structure that is often presented in mass media. The bully’s followers, as well as defenders, bystanders and witnesses, all played a significant role in the outcome of the story.
- Shape-shifting between roles frequently occurred. Defenders became victims, victims turned into observers, and bullies turned into the bullied.
- Adults played a particularly important role, as either heroes or villains. In contrast to the teens, the undergraduate stories did not include any adults. One of the undergraduates conceded that parents could have a role but only “from the outside”. This may reflect the undergraduates’ new autonomy from their own parents, or perhaps, an increased awareness of cyberbullying in the adults who surround today’s teens.
- The boundaries between what happens face-to-face and in social media were invisible. In these stories, the story stretched beyond the school and even the social media environment, into the personal lives of the victim (unlike in many bullying studies, where the story seems to end at the school yard gates).
- Within the shared narratives, as well as the ensuing discussion, participants indicated a wide range of values at play: establishing identity, responsibility and accountability, seeking help and support, empathy, social acceptance, justice and fairness, retribution, and the tension between individuality and conformity. Interface and interaction design, then, must take into account and seek to mediate a complex ecology of social values to have an effect on cyberbullying behaviors.
Conclusion

Narrative Inquiry is a form of storytelling and, as with many stories, it can be a conduit for lessons about life. Asked to tell a cyberbullying story, the participants uniformly presented one key message: Cyberbullying is a story of power and those who choose to defend others risk becoming the victim if they have no power themselves. Although the participants were asked to propose design interventions that would encourage reflective thinking, in actuality they went beyond this instruction to suggest functionality that would redress the balance of power, giving the victims of cyberbullying, as well as the circle of people who surround them, more leverage for solving the cyberbullying problem. These design interventions and others will be explored in future papers.

References


Paradigm Shift of LIS Education in the North East Region India

Bobby Phuritsabam  Th Purnima Devi  Memori Sagolsem
bobbyphster@gmail.com

Department of Library and Information Science, Manipur University

Abstract

This paper traces the emergence of library and information science education in North East India. Describe the current status, the different patterns, levels and modules of LIS education. Provide an overview of the programmes being offered by various universities in North East India. Discuss the needs for maintaining uniform model curriculum and standards in LIS education. Also discuss the trends, problems and challenges face by the LIS professionals. Discuss the approaches to prepare the LIS professionals to face the challenges of the future.

Keywords: LIS education, North East India, curricula, faculty strength, educational institution

Introduction

North-Eastern region of India commonly known as the “Seven Sister” includes the states. The region covering an area of 1,010,493.8 sq km with a total population of 35,623,086 is located at the corner most part of country with Assam the only state which links with the rest of the country by a narrow strip of land. This brings them closer to the world of Library and Information Science Education. North-Eastern region known for its geographical isolation, the region lends out a distinctive quality of life style and cultural to the rest of the country. Manipur for instance is popular for its dance and folk culture around the world.

Geography

North-Eastern part of India is a sensitive border zone where India meets for international countries – Bhutan, China, Myanmar and Bangladesh. Geographically even this remote section of the Indian land masses has elements of all three major physical features of sub-continents. The western mountains are a part of the Himalayas and the south of the Brahmaputra valley is part of the same fertile plains that make up the Ganges basin. The Khasi Hills, which lie further south an outcrop of the ancient peninsular rock of southern India.

Educational Institutions

Due to geographically constraints the region expect for Assam is of a recent phenomena in the field of education. Because of poor economical background the region had to depend more on central packages for setting up educational institution and its infrastructure. Despite the poor educational background the state governments are not giving top priority to education and its related field in budgetary allocation. Following are the institution/ universities selected for the proposed study. There are four universities selected on the basis of data that are collected during the course of study.

Political and Population of North East Region India

Assam: Assam has an area of 78,438 sq km, with its state capital at Dispur. The population of the state in 2001 is 2,66,38,407 with 23 districts. Manipur: Manipur has an area of 22,327 sq km with its state capital located at Imphal Comparising of 2,230 sq km surrounding by hills on all sides. In 2001, the population of the state stood at 23, 88,634.
Meghalaya: Meghalaya has an area of 22,284 sq km with a total of 7 districts is habited by a total population of 2,306,069 according to the 2001 census. Mizoram: Mizoram cover an area of 21,081 sq km with a total of 8 districts with its State capital at Aizawl is habited by a total population of 8, 91,058 according to the 2001 census.

Educational Institution of North East India

Assam: Assam has well established universities and Institution of good records for Library Educational point of view. There are five Universities. They are Gauhati, Jorhat, Silchar, Didrugurth and Tezpur Universities. Gauhati University conducts the course of Library and Information Science. The department was started as evening course in the year 1966 with University Librarian Dr Jogeswar Sarma as Head of the Department leading to Bachelor of Library Science (B.Lib Sc.) degree and in the year 1979 Master of Library Science (M.Lib.Sc.) course was introduced. In 1983 both B.Lib.Sc. and M.Lib.Sc. Courses become one year full time day courses, and in the year 1985 the course is restructured as Bachelor of Library and Information Science (B.L.I.Sc.) & Master of Library and Information Science (MLISc) course with some changes in the syllabus. Since 1987 the Ph.D. programme is introduced in the department. From the academic session 2001-02, as per UGC guide lines two years integrated MLISc course with four semesters has been introduced. Keeping in mind about the changing trend of the society the courses are again revised from this academic session 2007-08 more importance has been given to the use of IT in libraries. Candidate seeking admission to the course must posses the Bachelor's degree in any discipline recognized by the Gauhati University provided that such a candidate obtained a minimum of 45% marks in aggregate in Major subjects or 50% marks in aggregate in general course. The students are required to appear in an Admission Test conducted by the department and must have obtained a minimum 50% mark in the test and students are required to appear in a viva-voce Examination Test. Presently the department has three regular faculty full time and one guest faculty.

Manipur: The state has two Universities well established and catering the need of higher education. Manipur University was founded on 5th June 1980. It is located a Canchipur in the southern part of Imphal, the Capital. The state also has a reputation of having good Medical Institute like Regional Institute of Medical Science, having a well establish library with good collection handled by qualified professional. Another feature about the state is that the Central Agricultural University came in being in 1992. This university has a wider jurisdiction covering Sikkim. Manipur is one of the states in the region that came under library legislation. The department of Library and information science Education was established on 2nd April 1986 by Professor M.R. Khumbhar, the University librarian of Manipur University as its head of the department. The Bachelor of Library and Information Science was started since 1996. The department has started Master in Library and Information Science program for one academic in two semesters was started in the year 2004-05. Currently Master in Library and Information Science (Integrated Course), two year integrated course program for two academic years in four semester was started in the year 2005-06. The department has been providing facilities for Ph.D program since 1987. The department has so far produced 12 Ph.D degree holders.

Meghalaya: The State Meghalaya has many libraries of regional importance. Meghalaya has North East Hill University namely; one such is Regional Documentation and Information Centre (RDIC) catering the needs of users in the entire North-East region India. The Department of Library and Information Sciences was established in 1985 with a new approach to Library and Information Science education in the country. The Department introduced for the first time a two-year integrated programme leading to Master's degree in Library and Information Sciences in the country. The same pattern of education has now been recommended by the UGC under Model Curriculum in Library & Information Science in 2002. The University Library is being used as a laboratory for the students for practical training under the guidance of faculty members and practicing librarians. Thus, the theory and practice are given equal emphasis through this integrated approach. The Department has been regularly revising and updating the syllabus keeping with the rapid developments of information technology (IT) and its crucial role in the library and information profession. The last revision of the syllabus was made in 2003. Since 1986 nearly 350 students have already got MLISc degree from the Department and seven students have been awarded Ph.D. The department is offering the following courses: M. Lib Sc, M. Phil and Ph. D. The department has five regular faculties and one visiting fellow professor. Intake capacity of the department MLISc: 25 Students, PhD: 15 but it takes 3 PhD Scholars.
Mizoram: Many institutions for higher education were affiliated to North East Hill University (NEHU) at Shillong in Meghalaya. Mizoram has One University. The Department of Library & Information Science was established during the academic session 2002-03 and it was formally inaugurated by Prof. A.K.Sharma, Vice Chancellor of Mizoram University on 20.8.2002 at the Central Library of the University. The first intakes of the students were for the B.L.I.Sc. Course and which was later changed over to the Two Years Integrated M.L.I.Sc Course. The Department is also offering M.Phil program. Student intake capacity of the department MLIsC: 15 students.

Assam University: The Department of Library Science has been established in the year 2009. The Masters of Library and Information Science programme at Assam University, Silchar educates students in the principles, practices and ethics of library and information science, imbues them with a sense of service to diverse populations, and prepares them to be lifelong learners and active leaders in a rapidly changing information and knowledge society. The intake capacity of MLIsC is 25 students.

Library Education is in a great state of change now as any time since its introduction as part of higher education. Today library school are faced with two grave problems, one an internal one within the university and other a need to redefine their role vis-à-vis society as a whole in North East region India. In the last two decades new communication and computer-based technologies have produced profound changes in library education. These technologies has been felt in day-to-day internal library operation, in interlibrary network arrangement, in context of education for librarianship and finally, in the profession’s self-perception.

According to Bell, “where industrial society is based on machine technology, post-industrial society is shaped by an intellectual technology. . . . if capital and labour are the major structural features of industrial society, information and knowledge are those of post-industrial society”.

Academic Programmes

The education for librarianship in India has evolved against the complex background. It has, therefore taken much of it own character from the divergent nature of Indian education. Library education exhibits its own picture of diversity with courses in librarianship offered at a variety of levels and standards. Opportunities are better now, and have increased due to the facts that many states in the country implemented the library legislation. There is also a growth of special and industrial libraries and many more openings are available to the librarians in the multinational organization. Beside, the status improvements are also seen in the university and they are treated as par with other departments. Ranganathan Committee Report enumerated the following courses in Library Science (CDC Report, 2001).

- Certificate Course
- Diploma
- Post Graduate Course
- Bachelor's Degree Course
- Master Degree Course
- Doctorate Course

Presently six levels of course are conducted in India for LIS Course starting from certificate to Doctorate Degree. There is also change in nomenclature as well mode and schemes particularly at the Master Degree levels. Table below present the level of course conducted at India.

Teaching Faculty of North East Region India

The need for full time teachers was recognized by UGC Review Committee (1965) which said: “While we welcome a close association between the departments of library science and the library in a manner to be determined by the universities, we are not in favour of the present practice of employing part-time teacher in the departments of library science drawn from university libraries. This in our view is not conducive to efficiency . . . therefore it is not proper to continue any longer the practice of engaging university library staff to be part-time teachers in the departments of library science”.

The trend over the past decades or so has been gradually to employ full-time teachers as well as a department head in the library schools in North East region of India. Currently teaching faculties' strength of the departments in North East region of India is given in the Table below.
Table 1: Level of Courses

<table>
<thead>
<tr>
<th>Levels</th>
<th>Nomenclature</th>
<th>Degree</th>
<th>Minimum Qualification Requirement</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Certificate</td>
<td>i. Certificate in Library Science</td>
<td>CLISc</td>
<td>SSLC, PUC</td>
<td>3-9 Months</td>
</tr>
<tr>
<td>Diploma</td>
<td>i. Diploma in library Science</td>
<td>DLISc</td>
<td>SSLC</td>
<td>6M-2 Yrs</td>
</tr>
<tr>
<td></td>
<td>ii. JOC in Diploma in library Science</td>
<td>DLISc</td>
<td>SSLC</td>
<td>2 Years</td>
</tr>
<tr>
<td>Bachelor</td>
<td>i. Bachelor in Library and Information Science</td>
<td>BLISc</td>
<td>Degree</td>
<td>1 Year</td>
</tr>
<tr>
<td></td>
<td>ii. Bachelor in Arts (Hons) (Pass) (Optional)</td>
<td>BA</td>
<td>10+2</td>
<td>2-3 Years</td>
</tr>
<tr>
<td>PG Diploma</td>
<td>i. P-G Diploma in Information Technology</td>
<td>PGDIT</td>
<td>BLISc</td>
<td>1 Year</td>
</tr>
<tr>
<td></td>
<td>ii. P-G Diploma in Archives and Documentation Management</td>
<td>PGDADM</td>
<td>Degree</td>
<td></td>
</tr>
<tr>
<td></td>
<td>iii. Post MLISc Diploma in Library Automation</td>
<td>PGDLAN</td>
<td>MLISc</td>
<td>1 Year</td>
</tr>
<tr>
<td>Master</td>
<td>i. Master in Library and Information Science</td>
<td>MLISc</td>
<td>BLISc</td>
<td>1 Year</td>
</tr>
<tr>
<td></td>
<td>ii. Master in Library and Information Science (Integrated Course)</td>
<td>MLISc</td>
<td>Degree</td>
<td>1 Years</td>
</tr>
<tr>
<td>Research</td>
<td>i. Master of Philosophy</td>
<td>M. Phil</td>
<td>MLISc</td>
<td>1-2 Years</td>
</tr>
<tr>
<td></td>
<td>ii. Doctor of Philosophy</td>
<td>Ph D</td>
<td>MLISc</td>
<td>2-5 Years</td>
</tr>
</tbody>
</table>

Table 2: Faculties of the North East Region

<table>
<thead>
<tr>
<th>University</th>
<th>Year of Establishment of the department</th>
<th>Professor</th>
<th>Reader/Associate Professor</th>
<th>Lecturer/Assistant Professor</th>
<th>Guest Professor</th>
<th>Guest-Part-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gauhati</td>
<td>1966</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NEHU</td>
<td>1985</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Manipur</td>
<td>2nd April 1986</td>
<td>Nil</td>
<td>2</td>
<td>2</td>
<td>Nil</td>
<td>2</td>
</tr>
<tr>
<td>Mizoram</td>
<td>2002-2003</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>Nil</td>
<td>Nil</td>
</tr>
<tr>
<td>Assam University</td>
<td>2009</td>
<td>Nil</td>
<td>1</td>
<td>1</td>
<td>Nil</td>
<td>Nil</td>
</tr>
</tbody>
</table>

Table 3: Current Program

<table>
<thead>
<tr>
<th>Department</th>
<th>Under Graduate Diploma</th>
<th>BLISc</th>
<th>MLISc (Integrated Course)</th>
<th>M. Phil</th>
<th>Ph.D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gauhati University</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>NEHU University</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Manipur University</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Mizoram University</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Assam University</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Problems/Challenges of LIS Education in North East Region India

Information has become a commodity to be brought and sold. The provision of information per se came only with the establishment of special libraries and "information bureaus" within government and industry at a much later date. But in North East Region development of industries is very poor due to lack of communication and transportation. The LIS educational function has always remained strong and since most libraries have had limited resources and personnel, the usual pattern has been to guide a user to the sources of the information he or she needs and except in the case of ready-reference fact questions, not to supply that information per se. Many countries education for work in documentation centers has been completely separated from that for positions in traditional university. This is very much true in the North East region India. The library schools in the north east region India are thus finding it very hard to prepare students of special libraries due to lack of infrastructure. Furthermore, the introduction of much new technology has meant the downgrading of many traditional library tasks. As Allen Veaner has indicated, "The old jobs have gone to the support staff, whose activity covers a wide range of essential work. The work of inputting, coding, and verifying bibliographic data, selecting vendor, maintaining book fund accounts, claiming series, searching, establishing entries, filing, revising filing, proofreading, furnishing a first level of reference service, doing a good deal of cataloguing, and in some libraries, even revision of cataloguing has moved to support staff." It is not surprising that library schools in the region have sought to diversify their programs of study to enable their graduates to seek jobs in the "information" sector. Today schools have added some phrase including the word "information" to their names: at first the term "information science," more recently "information studies."

Lancaster has identified some obvious problems in curriculum development of Information Science in the developing countries. There includes

- Lack of resources;
- Greater bureaucracy;
- Low demand of information service;
- Traditional teaching methods-text bases and lecture method;
- Lack of standardization, co-operation and centralization;
- Significant differences existing among types of libraries;
- Diversity of culture and languages.

From the above reasons there seems that it is not easy to change from a more traditional curriculum to one that may be radically different. Lancaster, while giving example of India states that rather rigid examination procedure discourages change. But in the North East Region rigid examination procedure is adopting.

Accreditation

Accreditation is considered as cornerstone of self-regulation to achieve better and effective performance. Accreditation assures all the stakeholders, the community, academicians, students, employer and other organizations about the institution's clearly defined and educationally appropriate objectives and the required conditions for the achievement.

The need for quality assurance, evaluation and sustenance has initiated University Grants Commission (UGC) working towards quality enhancement since its inception. As a result, the National Assessment and Accreditation Council (NAAC) were established in 1994 to access and accredit institutions of higher learning in accordance with national & international norms and standards. Quality is emerging as a key value in all educational modes and methods in the processes of globalization of education. India has taken up the quality assurance and accreditation as means for reforming and upgrading standards in Indian system of education (Karisiddappa, 2005).

There are all India accreditation bodies which assess and certify institutions as a whole, individual department are not certified. There are no norms for the library schools except the general ones from the UGC (Satija, 2006). Graduate from North East Region India do not bear any stamp of quality assured. Not having a seal of any outstanding school with a name students, have to fend themselves in the job market. This way we are neither helping the students nor the employer.
Curriculum Areas

Revising the curriculum of schools of library and information science we need to look at what types of jobs we hope to prepare our students for. Most cite these areas as paramount: design of information system, management and marketing of information systems, creation of data banks and their marketing, and online searching.

Peter F.Oliva, in his book, Developing the Curriculum (1997), list these curriculum is statements. Curriculum is

- That which is taught in school;
- A set of subjects;
- A program of studies;
- A set of materials;
- A sequence of courses;
- A set of performance objectives;
- A course of study;
- Everything that goes on within the school, including extra-class activities, guidance, and interpersonal relationships;
- That which is taught both inside and outside of school directed by the school;
- Everything that is planned by school personnel;
- A series of experiences undergone by learners in school; and
- That which an individual learner experiences as a result of schooling (Nartarajan, 2006).

One emerging core area is “information Science”, “marketing” it tends to be a general introduction. The emphasis is on skill development in text processing, use of spreadsheets, electronic mail, and database management. Some assumptions for developing change in the curriculum:

Basic principles and concepts (i.e. content) can be specified;

- The particular manifestation of the content should be left to individual schools, i.e. whether required courses, series of elective, experiences, competencies, etc;
- The core content applied to all levels of program – undergraduate, postgraduate, doctorate – however, with different levels of emphasis;
- The core content applied with varying to all area of library/information science field.

Robert Taylor has stated, “The profession must recognize that is has an extraordinary set of skills. These skills, however, can only come to full fruition if they are separated from the library context. They are: the ability to design and to manage information systems in response to real needs; the ability to relate people to information and vice versa; the sense of service. These skills which, when married to a competent knowledge of information technologies, provide a base for the unique and central profession in the information society.” The core of an information program should cover the fundamental aspects of the basic functions . . . but should not be so extensive as to mitigate against flexibility in curriculum and program design. The content areas were organized into three categories: Knowledge, tools and skills.

Knowledge areas are philosophy and background, i.e. the foundations of information in society; environmental and contextual knowledge; and management knowledge.

Tools were identified as either quantitative/analytical, i.e. system analysis, research methods, logic, bibliographic control, abstracting and indexing, and data structures, collection development, technical services.

Skills requirements: communication, technological skills (programming, online searching, database management), and interpersonal skills.

Conclusion/ Suggestion

The future of LIS education is difficult to encapsulate early in the 21st Century. The departments hold, in common their teaching and research about organization, management, and use of information.

- North East Region LIS school strengths must increasingly reflect the unique composition of faculty and institutional alliance;
• There should be cooperation amongst the professional’s faculty staff & amongst the departments of the region;
• LIS educators must trend to be optimistic about their future;
• LIS school of the region greatest challenges are finding new faculty and new administrative leadership for their schools, and particularly leadership able to continue to adapt, take risk and assert leadership within their institutions and across the professions;
• Change is inevitable, but not all changes are equally possible or desirable;
• LIS education in North East Region of India is at the crossroads;
• LIS school of the region should concentrate exclusively on providing quality education and this will necessary in including library automation, networks, retrieval systems, and other information technology related courses.
• Graduates of these schools must be expected to work in libraries & other multinational companies.

The beginnings of this trend are expected to start. It is likely, LIS schools of the region will seek to expand their programs and attract more students by offering undergraduate courses in information technology. The intended trends are not meant to be a blue print for the future: but to stimulate discussion and to aid in planning a more responsive program for training and educating information professionals of the North East Region of India.

References


Using Digital Book Metrics for Navigation and Browsing

Michael Huggett
University of British Columbia
m.huggett@ubc.ca

Edie Rasmussen
University of British Columbia
edie.rasmussen@ubc.ca

Abstract

As the scholar’s work migrates from print to the digital realm, new ways of browsing, navigating and searching collections of digital books are needed. The Back-of-Book Index is a carefully crafted source of information on a book’s vocabulary and concepts, and if aggregated across multiple books, for a subject domain as a whole. Using a test collection of digital books in a variety of domains, we explore the use of index vocabulary to derive a series of metrics to indicate the relationships between index vocabulary and the digital collection, and the relationships between the books within a digital domain. We are investigating ways in which these metrics can be used to facilitate navigation and browsing at the domain level, to identify the most appropriate works within a digital collection for a particular subject or topic.

Keywords: digital books, back-of-book indexes, vocabulary, collections, browsing

Introduction

Millions of books are available in digital form as a result of mass digitization projects (such as the Gutenberg Project, the Million Books Project, the Open Content Alliance, and Google Books (Coyle, 2006)). As the scholar’s work migrates from print to the digital realm, new ways of browsing, navigating and searching collections of digital books are needed. Currently, digital collections are searched at a macro level through metadata such as author, title, subject, etc., and at a micro level through keywords. In contrast, the back-of-book index (BoBI) operates at an intermediate level based on significant terms and concepts identified by a human indexer. Though relatively unstudied, the BoBI is a traditional knowledge structure to support search and browsing in the print domain (Jörgensen & Liddy, 1996; Liddy & Jörgensen, 1993); a few studies have examined its use within a single digital text, usually in comparison to a search engine (Abdullah & Gibb, 2008; Chi et al., 2006; Egan et al., 1989; Liesaputra, Witten & Bainbridge, 2009).

Because the BoBI is carefully crafted by a human domain expert, it is a rich source of information on a book’s vocabulary and concepts, and if aggregated across multiple books, for a subject domain as a whole. The literature on BoBIs suggests that indexes serve a critical role in locating information in print and digital books, and that the indexing process generates a vocabulary that is richer, more structured, and more concentrated than that found in the book itself (Anderson & Pérez-Carballo, 2001; Gratch, Settel & Atherton, 1978). Such aggregate indexes could serve multiple purposes: to classify books based on summarized content (Enser, 1985), to generate a domain-descriptive vocabulary, to structure knowledge within the collection, and to navigate within the collection—applications which we are investigating in the Indexer’s Legacy project (Huggett & Rasmussen, 2012).

In this poster we describe the use of index vocabulary to derive a series of metrics to indicate the relationships between index vocabulary and the digital collection, and the relationships between the books within a digital domain. We are investigating ways in which these metrics can be used to facilitate navigation and browsing at the domain level, to identify the most appropriate works within a digital collection for a particular subject or topic.

Acknowledgements: Funding from the UBC Hampton Fund and the Social Science and Humanities Research Council is gratefully acknowledged. We also thank our Graduate Research Assistants who have sourced and pre-processed the many books for our test collections.


Copyright is held by the authors.
Building a Test Collection

Our test collection consists of seven domain corpora in the Arts (Art History, Music), Humanities (Economics, Cooking), and Sciences (Geology, Anatomy, Darwin). Each domain is comprised of over 100 publicly-available digital books. The index is extracted from each book’s PDF, and cleaned of OCR errors to leave only valid words. Each index entry is then ‘expanded’ to include on a single line its main heading, subheading(s) and page references. These expanded lines are then aggregated into a single file, sorted alphabetically, and compressed back into standard indented index format to create a single meta-index for an entire domain. The meta-index can be searched and browsed using the Meta-Dex User Interface¹ (Huggett & Rasmussen, 2012).

As the indexes are processed, we collect and calculate a variety of metrics that characterize the vocabulary, the books, and the domain. The metrics can be used to supplement the meta-index in browsing and navigating the digital collection.

Domain- and Vocabulary-Level Properties of the Test Collection

The basic properties of each domain are shown in Table 1. Each domain dominates in some measure. The domains show a high degree of variability in the number of tokens (i.e. instances of terms) and unique terms, within both content (i.e. the book’s chapters) and index. The table also shows high variability in the number of main entries (anchors), sub-entries (subs), and page references (refs).

Table 1. Domains Compared by Basic Properties

<table>
<thead>
<tr>
<th>domains</th>
<th>anatomy</th>
<th>art history</th>
<th>cookbooks</th>
<th>darwin</th>
<th>economics</th>
<th>geology</th>
<th>music</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>num files</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
</tr>
<tr>
<td>content terms</td>
<td>187366</td>
<td>165400</td>
<td>71519</td>
<td>152237</td>
<td>129897</td>
<td>194903</td>
<td>161315</td>
<td>151824</td>
</tr>
<tr>
<td>content tokens</td>
<td>385537</td>
<td>381590</td>
<td>326963</td>
<td>450974</td>
<td>551182</td>
<td>6537858</td>
<td>3927793</td>
<td>5152677</td>
</tr>
<tr>
<td>content tokens / term</td>
<td>46.1948</td>
<td>23.0712</td>
<td>45.7166</td>
<td>29.5970</td>
<td>42.4323</td>
<td>32.7263</td>
<td>24.3486</td>
<td>34.8095</td>
</tr>
<tr>
<td>index terms</td>
<td>27276</td>
<td>37962</td>
<td>9430</td>
<td>33290</td>
<td>15595</td>
<td>38980</td>
<td>31200</td>
<td>27676</td>
</tr>
<tr>
<td>index tokens</td>
<td>544537</td>
<td>233320</td>
<td>218963</td>
<td>210818</td>
<td>168526</td>
<td>276315</td>
<td>207960</td>
<td>265777</td>
</tr>
<tr>
<td>num anchors</td>
<td>16142</td>
<td>24222</td>
<td>4943</td>
<td>20241</td>
<td>8372</td>
<td>25723</td>
<td>18378</td>
<td>16860</td>
</tr>
<tr>
<td>num subs</td>
<td>130096</td>
<td>56671</td>
<td>56401</td>
<td>54074</td>
<td>44993</td>
<td>72105</td>
<td>49835</td>
<td>66394</td>
</tr>
<tr>
<td>subs / anchor</td>
<td>8.06</td>
<td>2.34</td>
<td>11.81</td>
<td>2.67</td>
<td>5.37</td>
<td>2.80</td>
<td>2.71</td>
<td>5.1086</td>
</tr>
<tr>
<td>num refs</td>
<td>226456</td>
<td>99152</td>
<td>78161</td>
<td>96944</td>
<td>46111</td>
<td>138335</td>
<td>98331</td>
<td>111927</td>
</tr>
<tr>
<td>refs / book</td>
<td>2220.16</td>
<td>972.08</td>
<td>766.28</td>
<td>950.43</td>
<td>452.07</td>
<td>1356.23</td>
<td>964.03</td>
<td>1097.3259</td>
</tr>
<tr>
<td>refs / anchor</td>
<td>14.03</td>
<td>4.09</td>
<td>15.81</td>
<td>4.79</td>
<td>5.51</td>
<td>5.38</td>
<td>5.35</td>
<td>7.8514</td>
</tr>
<tr>
<td>refs / sub</td>
<td>1.74</td>
<td>1.75</td>
<td>1.34</td>
<td>1.19</td>
<td>1.02</td>
<td>1.92</td>
<td>1.97</td>
<td>1.6471</td>
</tr>
<tr>
<td>shared anchors</td>
<td>6879</td>
<td>7328</td>
<td>1966</td>
<td>6906</td>
<td>2840</td>
<td>8775</td>
<td>6713</td>
<td>5915.2857</td>
</tr>
<tr>
<td>shared _%</td>
<td>42.62</td>
<td>30.25</td>
<td>39.77</td>
<td>34.12</td>
<td>33.92</td>
<td>34.11</td>
<td>36.53</td>
<td>35.9029</td>
</tr>
<tr>
<td>coherence _%</td>
<td>3.7020</td>
<td>1.5836</td>
<td>3.5531</td>
<td>1.8788</td>
<td>2.4332</td>
<td>2.0634</td>
<td>2.2559</td>
<td>2.4957</td>
</tr>
</tbody>
</table>

The table’s measures indicate some interesting properties. A low number of anchors suggests a ‘compactness’ of the domain’s set of concepts. A higher number of subs per anchor suggest that a domain’s individual concepts are more thoroughly elaborated, as does a higher ratio of references (per book, anchor, or sub).

Three calculated measures show subtler relationships:

- **Shared anchors** indicates the number of main entries in the domain that are shared between at least two books, providing a base measure of agreement of important concepts within the domain.
- **Shared _%** reflects the proportion of main entries in the domain that are shared between books. A higher value indicates better agreement on core concepts between the books of that domain. The

¹ http://meta-dex.no-ip.org

765
determination of whether low scores indicate a broad vocabulary or specialization into well-defined sub-areas may be solved by clustering books by dominant terms—a subject of future research.

- **Coherence, %** shows the degree to which all of the domain’s main entries appear in all of its books. If all of a domain’s entries appear in all of its books, the coherence is 100%. If none of books share any of their entries with another book, the coherence is 0%. The remarkably low degree of coherence of entries in each domain reveals how little texts agree on what is important, and how to express it. In other words, while a third or more of all main entries are shared within a domain, the vast majority are shared with very few other books. The degree of dispersal could possibly be lessened through better use of stemming algorithms that can aggregate strongly-related entries by variants such as singular and plural forms.

How much can we say about the domains based on these numbers, and to what degree of confidence? At this stage much is conjectural, and informed by our working familiarity with the domains. The numbers seem to support statements to the effect that Anatomy focuses on listing and organizing facts without much conjecture or argumentation, whereas Economics exposes ideas more by argumentation than by relying on a fixed set of discrete, specific facts. Clearly there is much more that we could do to compare domain properties—this forms a part of our future work.

**Book-Level Properties of the Test Collection**

The book-level properties across different domains (Table 2) mirror the domain-level properties, but the variability of size and formatting of each domain is more clearly exposed in the standard deviations (SD). For the number of anchors per book, the SD comes close to the average, whereas for subs and tokens the median is often lower and the SD often noticeably higher than the average, which suggests the influence that individual large books have upon the metrics of a domain.

<table>
<thead>
<tr>
<th>books</th>
<th>anatomy</th>
<th>art history</th>
<th>cookbooks</th>
<th>darwin</th>
<th>economics</th>
<th>geology</th>
<th>music</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>num books</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
<td>102</td>
</tr>
<tr>
<td>anchors avg</td>
<td>685.09</td>
<td>533.42</td>
<td>260.69</td>
<td>476.25</td>
<td>289.67</td>
<td>717.87</td>
<td>528.38</td>
<td>498.7671</td>
</tr>
<tr>
<td>anchors med</td>
<td>509.00</td>
<td>368.00</td>
<td>179.00</td>
<td>337.00</td>
<td>231.50</td>
<td>509.00</td>
<td>466.50</td>
<td>371.4286</td>
</tr>
<tr>
<td>anchors SD</td>
<td>537.03</td>
<td>463.42</td>
<td>276.95</td>
<td>404.13</td>
<td>214.77</td>
<td>641.72</td>
<td>337.56</td>
<td>410.7971</td>
</tr>
<tr>
<td>subs avg</td>
<td>1272.24</td>
<td>254.16</td>
<td>496.57</td>
<td>279.27</td>
<td>239.72</td>
<td>367.74</td>
<td>213.50</td>
<td>446.1857</td>
</tr>
<tr>
<td>subs med</td>
<td>696.00</td>
<td>101.50</td>
<td>234.50</td>
<td>160.50</td>
<td>138.50</td>
<td>184.50</td>
<td>142.00</td>
<td>236.7857</td>
</tr>
<tr>
<td>subs SD</td>
<td>1725.52</td>
<td>447.27</td>
<td>795.29</td>
<td>392.85</td>
<td>295.10</td>
<td>553.74</td>
<td>288.22</td>
<td>642.57</td>
</tr>
<tr>
<td>tokens avg</td>
<td>6609.31</td>
<td>2859.50</td>
<td>2461.57</td>
<td>2847.88</td>
<td>2255.41</td>
<td>3234.90</td>
<td>2508.73</td>
<td>3253.9</td>
</tr>
<tr>
<td>tokens med</td>
<td>3999.00</td>
<td>1319.50</td>
<td>1287.00</td>
<td>1827.50</td>
<td>1490.50</td>
<td>2274.00</td>
<td>1732.50</td>
<td>1990</td>
</tr>
<tr>
<td>tokens SD</td>
<td>8357.62</td>
<td>3841.09</td>
<td>3174.88</td>
<td>3895.45</td>
<td>2396.85</td>
<td>3058.93</td>
<td>2252.00</td>
<td>3853.8314</td>
</tr>
</tbody>
</table>

To get an idea of the properties of individual books in the test domains, we compared them by basic properties (Table 3). First we ranked the books by a simple count of their index entries, and by the proportion (p) of that count to the total number of unique entries in the domain as a whole. Across domains, we find that a handful of books hold a disproportionate number of the entries available in the domain. The proportion of usage in the top book ranges from 8% (Art History) up to 16% (Anatomy) of available entries.

(Left-to-right are tables for number of index entries, representativeness, and information gain per book. Book ID numbers are listed in the books column. Note the close agreement of book rank across tables.)

Second, we ranked the books by what we call their representativeness: the proportion of entries that they share with other books of the domain. Again, a small number of books share a disproportionate number of their entries: the top book in a domain ranges from 14% (Art History) up to 32% (Anatomy).

Third, we ranked books by their information gain, measured with respect to a book’s effect on domain coherence. We define coherence as the degree to which all of the books of a corpus share all of the available index entries in the corpus (see above). A book’s information-gain value is calculated as the effect on domain coherence if the book were removed from the domain. If a book shares entries with many other books, the book is effectively a ‘hub node’ in a shared-entry network model: removing a hub can drastically reduce a network’s connectedness, in the worst case potentially splitting the domain into discrete unconnected sub-networks.

We find that only a few books in each domain score highly as ‘hubs’, and that the level of gain tends to drop off rapidly. The gain of the top book between domains also varies from 4.45 (Art History) up to 13.03 (Anatomy), indicating that the top books in Anatomy capture more of the gist of their field than do the top books of Art History. In comparing these three measures, the chief observations are that books that have many entries also tend to share more of their entries with other books (representative), and tend to act as central hubs that bind the domain together (information gain).

It may seem self-evident that big books are important, but the results say more than that: a book with a lot of entries and a high information-gain value is broad and general, rather than highly specific and exhaustively detailed, since even a big book would have a low information-gain score if it shared few of its entries with other books.

Discussion

The main goal of the Indexer’s Legacy project is to explore ways in which the back-of-book index can contribute to the effective use of digital books within the subject domains of a collection. The metaindex provides a topic-based approach for searching a domain, but we are also interested in exploring domains based on their vocabulary, the unique properties of individual books, and relationships between books. We are currently building tools that employ visualizations and network models to support domain navigation. These tools address questions such as “Which books are most central to the domain?”;
“Which books have the greatest coverage?”; “What book is most closely related to another?” We are also interested in developing new metrics that will better support exploration at the domain and collection level.

**References**


The Usability Study on the Multicultural Children’s Book Project of the National Library for Children and Young Adults (NLCY) in Korea

Wooseob Jeong
Wij8612@uwm.edu
Renee Kapusniak
kapusni2@uwm.edu
Hyejung Han
hanh@uwm.edu
School of Information Studies, University of Wisconsin-Milwaukee

Abstract

In this study, usability issues at the multicultural children’s book project web site of the National Library for Children and Young Adults (NLCY) in Korea were identified, particularly for children users. Suggestions for improvements include brighter colors, bigger fonts and graphics, more lively animation and sound, easier navigation, consistent interface, summary availability, better organization of information and labeling, and multilingual searching and browsing function. A user study is being planned for future development.

Keywords: multicultural children, usability, interface design, National Library for Children and Young Adults

Introduction

Beginning in 2009, the National Library for Children and Young Adults (NLCY) in Korea has made about 300 children’s picture books published in Korea available on the web, in 6 different languages in the form of adopted animations accompanied by storytelling performances. This large project is one of many efforts by the Korean government to educate and assimilate foreign populations, along with their children, who have been arriving since the 1990’s in this still extremely homogeneous country.

A selection committee, consisting of the librarians at the NLCY and external experts, chose a list of quality books. The library contacted the publishers to clear copyright which were subsequently cleared with a modest fee. The ongoing project is currently available at the NLCY homepage at http://lscc.nlcy.go.kr:8000/. Figure 1 shows a screen shot of its front page.

Purpose of Study

In spite of the massive quantity of translated, animated and performed storytelling in this project, the interface of the web site is not ideal for users, particularly for children. In this study, we would like to identify usability issues on this site based on findings from the related literature and provide suggestions for future improvement to create a better interface.

Literature Review: Children’s Information Seeking Behavior

In order to learn how children’s information seeking behavior influences children’s interface design, various research studies about children’s information seeking behavior were reviewed. Dresang (2005) points out that children’s information seeking behavior related to children’s digital media use might be that the young are missing rich information from the environment due to their information seeking skills not being fully developed. Bilal (2005) tells that 43% of children felt frustrated during the search process because of the zero hits of children’s retrieval using Yahooligans! In addition, 85% of children were also thank our Graduate Research Assistants who have sourced and pre-processed the many books for our test collections. Jeong, W., Kapusniak, R., & Han, H. (2013), The Usability Study on the Multicultural Children’s Book Project of the National Library for Children and Young Adults (NLCY) in Korea. iConference 2013 Proceedings (pp. 769-772). doi:10.9776/13371

Copyright is held by the authors.
motivated to use the Web for increasing self-confidence, challenge, discovery and for convenience. According to the research of Madden, Ford, Miller and Levy (2005), the factors which determined children’s ability to search successfully were the amount of experience the children had of using the Internet, the amount of guidance from adults and friends (Druin, Foss, Hutchinson, Golub, & Hatley, 2010) plus, the ability of them to explore the virtual environment and to use the tools available. Lu (2010) points out that children’s information need should connect to children’s other needs at different developmental stages. In a recent study, Bilal & Sarangthem (2009) stated, “children information seeking was characterized by seven modes of behavior: Start, Recognize, Browse, Differentiate, Read, Explore and Finish.”

Preliminary Findings

Colors, Fonts, and Graphics

The NLCY (URL: http://lscc.nlcy.go.kr:8000/index.php?nFlag=Y&sctnation=2&&nation=) has a background which currently has some basic colors in muted tones with small fonts displaying the title of the story underneath a small graphic depiction of a page in the story. Children prefer to have a portal page which has bold, colorful backgrounds and graphics which include fonts that are large; such as a 14 pt font for young users and 12 pt font for older users (Nielsen, 2010) with suitable vocabulary for their age (Bilal & Bachir, 2007). Creating a page with brighter colors, larger graphics and fonts will enhance the page and attract more users. It is difficult to know what the languages are on the NLCY interface and the flags depicted show ambiguous identification of each country. Large, Beheshti and Rahman (2002) suggested the use of attractive screen designs based on color, graphics and animation, plus keyword search and browseable subject categories for the design of children’s web portals.

Animation and Sound

Animation and sound are important for children to like and want to interact with the interface page (Nielsen, 2010). Druin (2005) recommended an interface which can be modified to increase children’s retrieval success for digital libraries. The NLCY has animation and sound, however there are some limitations. Many scenes only have a character’s head or one object moving while the remaining screen is static. In other words, only a small part of the screen is animated, which may make the children watching it bored. In regards to sound, there is only one storyteller on the site with very limited sound effects. The stories need more animation and sound to create more interest and engagement with the story.

Difficulty in Navigation

To navigate to the next page of stories on the interface, a child needs to click at the bottom of the page, a small numeral or a tiny red arrow. It is difficult for children to find and accurately select a small icon to advance to another page which can cause the child to become discouraged in navigating the site (Hutchinson, Druin, Bederson, Reuter, Rose & Weeks, 2005). Children should have buttons or larger targets on the interface; therefore the child can easily move and accurately navigate to the stories.

Inconsistent Interfaces

Within each story, there are small interactive buttons which hide or allow the speaker of the story or the written words of the story to appear while the user is listening to the story. It was found in the course of this study, these buttons varied in location from one story to the next. With some stories, the buttons are located on the left side margin of the book and in other stories; the buttons were located on the right side of the book. The buttons also changed color along with being housed in a one button icon or they had their own separate button. Inconsistencies in the layout of a user interface will affect how the user feels about the site and whether they want to continue using it (Ahmed, McKnight & Oppenheim, 2006). Creating a layout which follows the same type of format for each storybook will increase the satisfaction of the users.
Summary Availability

Hutchinson and others (2005) explained how older children rely on book summaries when selecting a story to read. The NLCY provides some summaries for their books on the interface; however, they are inconsistent throughout the selections with some books providing a summary, while others do not. Having an overview of a story might aid a user in making a choice in selecting the book to read (Bederson, Quinn, & Druin, 2009). Without the summary, the user will either choose not to read it or they have to enter and start reading the story to determine the choice.

Information Organization and Labeling

Labeling and organization of information is important, especially for children due to their cognitive and physical development. Within the NLCY, the stories are organized with nine books displayed on a page with the titles of the stories underneath each pictorial representation of the book. Children typically choose books by category or from a physical characteristic of the book. Having individual categories which represent the type of books within the collection will aid the users in the process of browsing or searching for a book they would like to read (Bilal, 2002; Hutchinson et al., 2005; Reuter & Druin, 2004). Moreover, as part of the labeling on the interface, HELP is an essential component, especially on a children's interface page. Children like to have HELP that provides information which aids in the searching process (Bilal, 2002) The NLCY English page provides HELP, however it is labeled “Details”. Besides how some young children might not understand the definition of the word “Details”, the section itself only provides labeling of each facet on each specific interface page. Additionally, the section does not explain how to conduct a search on the site. A suggestion would be to provide “How To” pages to make it easier for children to search for stories on the site with perhaps screenshots or tutorials of navigating the pages.

Searching and Browsing

An interface should have more than one type of searching style available to accommodate all age levels and searching modes (Reuter & Druin, 2004). On the NLCY, searching for a story can be done via a search box, though the search box only allows for the Korean language. Being a multicultural storytelling site with multiple languages, the interface should be multicultural and provide for searches with other represented languages. Plus, the site only allows for a keyword search and according to Bilal (2002) who investigated children using a web search engine, children didn’t like using only a keyword search box, they would rather use an icon/metaphor (Druin, 2005) and other ways to search on the site (Bilal, 2002). Furthermore, icons/metaphors are needed to aid in the searching process which would aid a searcher who has difficulty spelling (Bilal & Bachir, 2007; Hutchinson, Rose, Bederson, Weeks & Druin, 2005). Icons/metaphors are usually objects which aid a child in the searching process by providing a recognizable item (Large, Beheshti, & Cole, 2002). Providing representative icons will allow users who are younger to find a story they would like without having to worry about interpreting the text presented on the page. Additionally, when searching or browsing, children typically prefer a faster load time to keep their interest (Bilal, 2002). The NLCY is inconsistent in the load time with some individual stories loading slower than others.

Conclusion

There is a great value in making 300 children’s books available to the public at the NLCY digital library in 6 different languages with their original pictures animated and their stories performed in a storytelling mode. However, to provide more pleasant access to the valuable site, particularly for children users, there is much room for improvement in terms of the user interface. Based on these findings, a user study is being planned for future development, which should be more intuitive for users, particularly children.
References


Examining Transdisciplinarity in iSchools Dissertations (2010-2011)

Kathryn Masten
University of North Texas
Kathryn.Masten@unt.edu

Abstract

Two defining characteristics of the iSchools are that they are interdisciplinary and aim to address societal problems (Dillon & Rice-Lively, 2006; Thomas, von Dran, & Sawyer, 2006). This study examines what “interdisciplinary” means in the iSchools context and the extent to which “societal problems” are being addressed through iSchools dissertation research. With this emphasis on broader societal impacts, the vision for the iSchools closely aligns with the concept of transdisciplinarity (Thompson Klein, 2008). Drawing from the literature about the continuum of cross-disciplinary approaches (unidisciplinary, multidisciplinary, interdisciplinary, and transdisciplinary), this study adapts and extends a transdisciplinary dissertation assessment tool created by Mitrany and Stokols (2005). This poster presents preliminary results from an assessment of 175 dissertations completed in 2010 and 2011 by students in the 21 iSchools as of 2009 (according to Wiggins and Sawyer, 2011). The purpose is to establish a snapshot of the level of transdisciplinarity among the iSchools.

Keywords: transdisciplinarity, interdisciplinary, broader impacts, iSchool, dissertation

References


----------------------------------


doi:10.9776/13376

Copyright is held by the author.
Abstract

The researchers propose to share first year results from VIEWS2, Valuable Initiatives in Early Learning that Work Successfully, a 3-year Institute for Museum and Library Service (IMLS) National Leadership Research Grant. Year 1 of VIEWS2, using a pre-post text experimental design was proposed to measure the early literacy outcomes for children birth to kindergarten who attend public library programs. Results stem from data collected by 20 iSchool student and faculty researchers in 40 libraries at 120 storytimes attended by 1440 children using innovative researcher-developed tools designed to observe children's behavior and adult program presentation. High positive correlation occurred between adult practices and developmentally appropriate children's outcomes at the group level. Some assumed best practices are in place, but some other research-based early literacy principles most important for children to learn to read successfully are relatively rare.

Keywords: public libraries, early literacy, IMLS National Leadership Research Grant, storytimes, informal learning environments

Introduction

From scrolls to codices to digital interactive eBooks, information leaders investigate ways to connect people to information through technology. While the written word continues to be the primary dissemination technological tool, illiteracy still separates people from information that can affect the quality of life. Early literacy storytimes in informal learning environments are one venue providing opportunities to change lives through literacy skill development. Central to early literacy is the idea that important reading skills need to be developed before the onset of formal literacy instruction (Sulzby & Teale, 1991; Whitehurst & Lonigan, 1998). Of these primary early literacy skills, alphabet knowledge, phonological awareness, and oral language have repeatedly been shown to have predictive validity for later reading proficiency (National Early Literacy Panel, 2008; Storch & Whitehurst, 2002; Whitehurst & Lonigan, 1998). The development of these preschool skills has also been shown to have associated economic benefits, such as increased maternal earnings and increased lifetime earnings (Barnett & Masse, 2007). Public libraries and other informal learning environments have committed resources to improving early literacy opportunities for children; e.g., through staff training and enhanced storytime offerings incorporating research-based primary early literacy skills.

Overview of Research

The widespread national commitment of libraries to early literacy programs as well as the lack of research on the impact of these programs informed the development of VIEWS2, Valuable Initiatives in Early Learning that Work Successfully. A thorough literature search yielded no research project assessing the impact of early literacy storytimes in public libraries despite their ubiquity. This 3-year Institute for
Museum and Library Service (IMLS) National Leadership Research Grant (October 2011 – September 2014) designed to find valid and reliable means to measure the early literacy outcomes for children birth to kindergarten who attend public library programs seeks to leverage public library and school partnerships to improve early literacy practices. The researchers involved in this grant propose to share initial VIEWS2 data and innovative approaches to investigating and evaluating early literacy within informal learning environments.

Research Design and Methodology

The design for Phase 1 and 2 of the research is a pre-post test quasi-experiment one, also not found anywhere in library research literature. Phase 1 collection included 120 storytime observations attended by 1440 children in 40 randomly selected Washington libraries (13 large, 13 medium, and 14 small). As can be seen from the map below, the 40 libraries in which the research took place are in virtually all geographic locations with the higher concentration where the population is the most dense yet are in many areas of the state. Three storytime types (Birth-18 months, 18-36 months, and 36-60-months) were observed. The children’s reactions and librarian’s delivery of each storytime were video-recorded and coded from the videos. The researchers developed a seminar that was delivered in winter 2012 to the graduate students that were collecting data and coding. Established validity and reliably trained 19 graduate students to use the BCPAF and PET tools. The data collection and coding occurred in spring 2012.

![Map of Washington showing the locations of the 40 libraries.](image)

*Figure 1. Locations of the 40 libraries.*

Research Tools

Previously established protocols for a tool, Benchmarks Curricular Planning and Assessment Framework (BCPAF) (Feldman, 2011), developed to observe children at the group level and based on Washington State Department of Early Learning Guidelines, were used in Phase 1. A companion Program Evaluation Tool (PET) tool was developed by the researchers to assess the program delivered by library staff was also used. Both tools were based on the premise of early literacy inclusion in storytimes. The coded data from BCPAF and PET were analyzed using SPSS descriptive and correlational results to uncover relationships between program delivery and children’s behavior. One of
the major findings from Phase 1 was the high positive correlation between the adult practices and developmentally appropriate children's outcomes at the group level. An addition finding was that some of the assumed best practices of librarians are in place in almost all of the storytimes, but there are some gaps to be filled, especially on some of the research-based principles that are most important for children to learn to read when they enter school.

**Ongoing Work**

During the second phase of VIEWS2, two additional tools will be used: the EL-Capstone instrument that summarizes adult early literacy concept awareness levels (Capps, 2011) and the Early Literacy Skills Assessment (ELSA) process developed by Highsmith. ELSA will be used in public libraries for the first time by VIEWS2 (See Campana & Dresang, 2012 for use of ELSA in home-based daycares). In line with the pre-post experimental design, the researchers will randomly select a group of the 20 of the 40 libraries and deliver an intervention to be based on the findings in Phase 1. Another cohort of graduate students will be reliably trained on using the instruments in the Winter 2013. The following Spring, the data collection process will begin again following the same process as in year one. This will help answer the research question of whether storytimes can be improved based on the information collected in Phase 1. Finally for the third phase the focus will be on how the grant will increase the early literacy link between school and public library.

**References**


Scientific Metadata Quality Enhancement for Scholarly Publications

Chun Guo  
School of Library and Information Science  
Indiana University Bloomington  
chunguo@indiana.edu

Jinsong Zhang  
Dalian Maritime University  
Dalian, China  
zjs.dlm@gmail.com

Xiaozhong Liu  
School of Library and Information Science  
Indiana University Bloomington  
liu237@indiana.edu

Abstract

Keyword metadata is very important to the access, retrieval, and management of scientific publications. However, author-assigned keywords are not always readily available in digital repositories. In this study, in order to enhance metadata quality, we explore different automatic methods to infer keywords from scholarly articles, including supervised topic modeling, language model, and mutual information. Evaluation results showed that the linear combination of mutual information and topic modeling with full text outperform other methods on MAP, while language model with abstract performed better than other methods on the measure of precision@10.

Keywords: keyword inference, topic modeling, language model, mutual information

Introduction

Keyword metadata is a very important access point for digital libraries. It provides a brief summary of the topics discussed in an academic publication. Although it is usually strongly recommended or required that an author provide keywords when they submit a paper for review, only a small number of publications have author-assigned keywords.

Take the distribution of publications with a certain number of keywords in the ACM digital library as an example (shown in Figure 1). For the 248,893 publications within the database, the number of keywords for a single publication ranges between 0-49. The average number of keywords per publication is 2.11. More than half of the publications (52.3%) do not have any author-assigned keywords, while 4.9% of them only have one or two keywords.

![Figure 1. Distribution of publications with a certain number of keywords in the ACM digital library.](image-url)


Copyright is held by the authors.
It is hard to ensure the quality of keyword metadata, since little guidance is provided to the authors when they assign keywords to their submissions. Although some publishing agencies have domain experts manually assign keywords for publications, it is very laborious work and can hardly be applied to very large collections. In this study, we explore inferring keywords for scholarly publications utilizing different automatic approaches.

**Literature Review**

Previous studies used two methods to infer keywords for documents: keyword assignment and keyword extraction. Keyword assignment, a.k.a. text categorization (Dumais et al., 1998), assumes that all potential keywords come from a predefined controlled vocabulary. Then machine learning model is trained to classify publications into those categories, while each publication is assigned with one or more category labels. However, this approach ignores the dynamic nature of author keywords. For instance, new concepts emerges everyday, a static controlled vocabulary is immediately out of date the moment it is created. Meanwhile, controlled vocabularies are usually created by domain experts, which, in most cases, may not reflect the interests of authors and readers.

The other approach for keyword inference is keyword extraction. It is not restricted to a set of candidate keywords from a selected vocabulary. Instead, any phrase in a new document can be extracted as a keyword. Tomokiyo and Hurst (2003) used language model to extract keywords from newsgroups. Kea (Frank et al., 1999) uses TFIDF (term frequency ∗ inverse document frequency) and normalized word position as machine learning features to extract keywords from documents. While extraction-based methods generate a more diverse set of keywords, some of the keywords extracted are not reasonable from a human perspective (Witten, Paynter, Frank, Gutwin, & Nevill-Manning, 1999).

In this study, we use author assigned keywords as predefined labels for keyword inference, and each keyword is represented by a topic model or a language model. We used keywords that frequently appear within a domain, which makes them more author/user centric.

**Method**

**Keyword Inference**

**LLDA.** Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is a widely used method for topic modeling. Labeled LDA (Ramage, Hall, Nallapati & Manning, 2009) further constrains LDA by defining a one-to-one correspondence between LDA's latent topics and user labels. Given a set of labeled documents, machine can learn the word-topic probability distribution, while each topic is represented by a label. As a result, unlike classical LDA, it is not necessary to interpret each topic or to set empirical topic number for a model. If documents used in the training process are well representative of the topic domain, then we can use the derived model to predict labels for other documents from the same domain.

**Language model.** Language Model is a topic dependent Information Retrieval Model. A document is treated as if it is generated by an unknown language model and we can estimate that model using statistical methods. If a document is relevant to a query, they are likely to be generated from similar language models. Based on this assumption, documents are ranked based on the likelihood of their background language model generating the given user query.

Vice versa, if we have a list of potential keywords for a topic domain, we can predict a publication's matching keywords by ranking all the potential keywords based on their likelihood of being generated by the background language model of the given publication.

**Mutual information.** Some publication venues require authors to provide category information for their paper during the submission process. For example, most ACM submissions require authors to pick categories from a controlled category list. Since both category words and keywords provide topical information, category words and keywords related to the same topic would highly likely co-occur in the same publication. Therefore, we hypothesize that combining mutual information with LLDA would further improve LLDA’s performance. We calculate mutual information score (MI_score) for each category-keyword pairs appearing in the corpus. Then we derive a new ranking score using a linear combination of mutual information score and LLDA score (LLDA_score) based on the formula below:
\[ \text{new score} = \alpha \cdot \text{MI\_score} + (1 - \alpha) \cdot \text{LLDA\_score} \]

\( \alpha \) is the parameter used for linear combination and different \( \alpha \) values were tested to achieve best performance.

**Evaluation**

The inferred keywords for each publication are evaluated against the original author assigned keywords. We use two indicators to measure keyword inference performance: mean average precision (MAP), and precision at 10.

We use greedy matching as the baseline method in the evaluation. We search each potential keyword from the full text of that article by using greedy matching. For example, if “music information retrieval” existed in the title, we wouldn’t use the keyword “information retrieval”. Matched keywords are ranked based on the position of their initial appearance in the article’s full text.

**Experiment**

**Data**

We used 41,370 publications from 111 journals and 1,442 conference/workshop proceedings on computer science (mainly from the ACM digital library) for the experiment, for which full text was extracted from the PDF files. The selected papers were published between 1951 and 2011. From these we extracted 20,394 publications’ text (accounting for 67.7% of all the sampled publications), including titles, abstracts, and full text. For the other publications, we used the title, abstract, and information from a metadata repository to represent the content of the paper.

**LLDA Model Training**

We sampled 20,394 publications (with full text) to train the LLDA topic model. Author-provided keywords were used as topic labels. For instance, if a paper has 6 author-provided keywords, our LLDA training would have assumed that this paper is a multinomial distribution over these 6 topics. During pre-processing we also clustered similar keywords if the edit distance between them were very small, e.g., “k-means” and “k means”, or if two keywords shared the same stemmed root, e.g., “web searches” and “web searching”.

Finally, we trained a LLDA model with 1,239 topic labels (keywords). These topic labels were used as potential keywords to be assigned to publications.

**Experimental Results**

<table>
<thead>
<tr>
<th>MAP</th>
<th>llda_abst</th>
<th>llda_full</th>
<th>lm_abstr</th>
<th>lm_fullt</th>
<th>greedy</th>
<th>mi_llda_f</th>
<th>mi_llda_f</th>
<th>mi_llda_f</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ract</td>
<td>ext</td>
<td>act</td>
<td>xt</td>
<td>matching</td>
<td>ull_0.2</td>
<td>ull_0.5</td>
<td>ull_0.7</td>
</tr>
<tr>
<td>0.1877</td>
<td>0.2632</td>
<td>0.2372</td>
<td>0.2133</td>
<td>0.2195</td>
<td>0.2714</td>
<td>0.2632</td>
<td>0.2632</td>
<td>0.2632</td>
</tr>
<tr>
<td>0.104</td>
<td>0.1453</td>
<td><strong>0.1784</strong></td>
<td>0.166</td>
<td>0.1573</td>
<td>0.1457</td>
<td>0.1453</td>
<td>0.1453</td>
<td>0.1453</td>
</tr>
</tbody>
</table>

The best performing method on MAP is the combination of mutual information and LLDA with full text (mi_llda_full_0.2). The best performing method on precision@10 is language model with abstract. Another interesting finding is that LLDA performs better with full text, while language model performs better with abstract, on both MAP and precision@10. It indicates that all terms within a document are important for LLDA inference, whereas language model favors terms that provide strong topical information. Apparently, paper full text contains more words than abstract, which helps the LLDA model to make more accurate inference. But it also brings more noisy terms, which hampers the performance of
language model. One the other hand, abstract, as a succinct summary of the publication, has a high concentration of topic terms, and works better for language model based keyword inference.

**Conclusion**

Keyword metadata is an important access point for publications in a digital library. However, most articles do not have any keywords assigned by their authors or only have one or two. Besides, manually assigning keywords to publications is a tedious work. In this paper, we explored automatic keyword inference using topic modeling techniques and full text data. The linear combination of mutual information and LLDA with full text outperformed other methods on MAP, while language model with abstract performed better than other methods on precision@10.

**References**


Teaching Tools, Applications, and Infrastructure for Digital Curation Through the Use of a Virtual Lab

Joseph Helsing  
Department of Computer Science  
University of North Texas  
Joseph.Helsing@unt.edu

Paulette Lewis  
College of Information  
University of North Texas  
paulette.lewis@unt.edu

Edward Warga  
College of Information  
University of North Texas  
edward.warga@unt.edu

Abstract

With an increased demand for digital curation, data management, archiving, and preservation the library and information science community has begun offering new education and training in these fields. A trend in this education is the use of online virtual labs to offer students hands-on experience with the tools and methodologies inherent to these fields. At the University of North Texas, we are also developing a four course education program in digital curation and data management including a course called Tools, Applications, and Infrastructure for Digital Curation that employs a virtual lab component. This paper discusses the first offering of the course; several challenges we faced; and the new strategies developed to address these issues and improve the course and virtual lab for the next offering.

Keywords: virtual lab, digital curation, elearning, online course

Introduction and Project Goals

Libraries have new opportunities to meet the needs for data management, sharing, archiving, and digital curation through new roles and services (Walters and Skinner, 2011). The library and information science community is responding by offering training and education programs focused on digital curation, data management, archiving, and preservation (Fulton, Botticelli, and Bradley, 2011; Harvey, and Bastian, 2012; Yakel, Conway, Hedstrom, and Wallace, 2011). A common trait among these academic programs is the development of virtual labs in which students can gain hands-on experience with the technology infrastructure and tools of the curation environment while working through problems and scenarios concerning digital content management. To work in an information technology rich environment, these professionals need practical knowledge and skills with the tools and processes used in digital curation. Such technology competencies are described by Lee (2009) in the Matrix of Digital Curation Knowledge and Competencies. The virtual lab component of these academic programs seeks to support students as they develop these competencies.

The University of North Texas received a grant from the Institute of Museum and Library Services to create a four course, competency-based curriculum in digital curation and data management (Moen, et al., 2012). The Information: Curate, Archive, Manage, Preserve (iCAMP) Project is a three-year curriculum development project to meet the need for digital curation professionals in libraries and other information centers. During the first year of the iCAMP Project, two courses were designed and implemented: Digital Curation and Data Management Fundamentals and Tools, Applications, and Infrastructure for Digital Curation. These courses are intended to give students an introduction to the concepts and technologies used in digital curation and data management.

As the second of four courses, Tools, Applications, and Infrastructure for Digital Curation offers an introduction to tools and processes used to conduct digital curation activities. The iCAMP Virtual Lab was developed to provide an environment for students to perform activities and gain experience with the tools, applications, and processes of digital curation. The Virtual Lab is a network of technologies including Linux environments pre-loaded with tools; VirtualBox files loaded with applications such as...
DSpace, EPrints, Omeka, and Archivematica; and a shared Fedora repository using a Drupal Islandora front end.

We first implemented the course in Summer 2012, and collected feedback from students using surveys and a focus group session. The data collected inform our course revisions during Fall 2012. Course revisions focus on improving the Virtual Lab technology infrastructure, the learning materials, and the activities of the course, with the goal of improving student performance, learning, and satisfaction. The revised course will be implemented and evaluated again in Spring 2013.

**Virtual Lab Infrastructure and Tools**

To enable students' hands-on experience with the technologies for digital curation in an online course format, we built a robust cyberinfrastructure environment. This included two primary areas for knowledge development: basic Linux and server knowledge; and digital curation tools and systems knowledge. We used two different systems to cover these areas of knowledge: VirtualBox and an Ubuntu Server.

The first system implemented used the VirtualBox by Oracle. We assumed that some students would be novices in the use of the console in the Linux environment, and we wanted to provide them a safe, destructible system. VirtualBox allowed us to make pre-setup Ubuntu installations that could be redownloaded if someone accidentally destroyed their home directory or an assignment. Students first downloaded the emulator, and then an Ubuntu VirtualBox preloaded with tools and files. Assignments required students to interact with the Linux environment and learn how to use the basic commands and operations to complete certain tasks.

Once students had gained sufficient skill with Linux, they started working on a local server installed and configured for the course. This server was similar to their VirtualBox but had a Fedora repository attached to a Drupal site via Islandora. On the server, we provided tools such as BagIt, ImageMagik, and a Fedora repository. Students performed activities informed by the OAIS reference model and the Digital Curation Centre Lifecycle Model. They engaged in these activities through guided exercises, using the tools in a manner similar to a professional digital curator. Students also performed activities with other digital curation software platforms including Archivematica, EPrints, and DSpace available to students as VirtualBox files.

Since all students were working on the same server for a portion of the course activities, it was very important to provide security and privacy for each student's files. Each student was chrooted into their home directory and only allowed sym links to necessary folders. This, along with strict permission settings, allowed students to work on the same server, while preventing cheating and potential destruction of other students' property.

**Technology Challenges, Student Evaluation, and Improvements**

Students expressed overall satisfaction and appreciation for the hands-on experience with the tools and technologies presented in the course and through the Virtual Lab. However, student evaluations and feedback pointed out three major areas that presented problems which we are now addressing in the revision of the course and technologies. First, some students realized that they lacked the technical skills necessary for full and ready participation in the course. Second, students reported that it took too long installing and configuring the technology used in the course. Third, students expressed the view that there was a lack of resources and materials available to help solve technology issues in the Virtual Lab, thereby causing delays in learning.

**Lack of Technical Skills**

Students expressed they lacked overall technical skills necessary for engagement in an information technology rich environment. Some students were unfamiliar with file transfer protocols and related software applications. Additionally, students had little to no fundamental knowledge of server administration and command line interfaces. Students also indicated the need for a Linux primer before beginning the course. To address this issue, we are creating a scaffolding course in Linux to provide the students with basic command line skills. We will also outline and publish the prerequisite skills and knowledge required for success in the course.
Time Installing, Configuring VirtualBox

The use of VirtualBox and appliances served the purpose of providing a safe environment for student exploration. However, this technology proved ineffective due to our inability to ensure a controlled laboratory environment for student practice. The variability among operating systems, memory, RAM and other software in the students’ computing environment proved to be a consistent challenge. Students expressed the need for more extensive assistance with the initial technology setup on their individual computers. Not only was it time consuming to download the VirtualBox and several appliances, students also reported that they spent too much time configuring hardware and software for the course. We provided installation tutorials for these tools, but some students encountered technical issues that we did not anticipate. For example, the automatic configuration in the Omeka appliance failed and students were not sure how to manually configure the appliance. Students also encountered problems using SSH/SFTP to transfer files into and out of the Ubuntu virtual machine. Students experienced numerous connection errors because the VirtualBox was unable to accept outside network connections. A considerable amount of time was spent testing various network configuration schemes to determine the correct settings needed.

Since several of our students had problems with using VirtualBox, we decided to forego the use of VirtualBox. Students will have access to an Ubuntu server to learn the basics about Unix environments. Additionally, instead of providing software like Archivematica and EPrints as a VirtualBox, these programs will be hosted on the server and accessed through a web interface. This will hopefully remove additional complications added by using a server emulator and give us greater abilities to diagnose and solve problems.

Lack of Resources and Materials

Students reported that we did not provide appropriate resources and materials to help them solve technology issues in the Virtual Lab. Students communicated they had to troubleshoot various technology errors and software conflicts throughout the course. This hampered the students’ abilities to complete assignments and projects and also caused delays in their learning. Students expressed a need for dedicated technical support personnel to help troubleshoot problems throughout the course. Our revised course will provide a knowledge base and helpdesk to support students with technology problems. We will work to identify common technology issues, gather more publicly available online tutorials, and create better documentation for existing infrastructure.

Conclusion

The iCAMP Virtual Lab is our first step toward building successful infrastructure to support this course. Our goals are to have students learn digital curation tools and methods in an environment modeled on the OAIS reference model and the DCC Digital Curation Lifecycle Model and develop the practical technology competencies required in the field of digital curation and data management. By listening to student feedback and responding with revisions to the course, the iCAMP team will refine the technology infrastructure and learning materials to provide an improved student learning experience in future implementations of the Tools, Applications, and Infrastructure for Digital Curation course.

References


A Prototype System for Heterogeneous Data Management and Medical Devices Integration in Trauma Resuscitation

Zhan Zhang
College of IST
Drexel University
zz87@drexel.edu

Aleksandra Sarcevic
College of IST
Drexel University
aleksarc@drexel.edu

Yuan An
College of IST
Drexel University
yuan.an@drexel.edu

Abstract

We propose a system for information acquisition, integration and presentation to support situation awareness during trauma resuscitation. The system consists of (a) medical devices for data acquisition, (b) a database and middleware applications for storing and processing data, and (c) a computer coupled with a large wall display for data presentation. We discuss the initial design of this system and the planned evaluation and implementation steps in the future.

Keywords: heterogeneous data, system architecture, healthcare, trauma resuscitation

Introduction

Trauma resuscitation is a safety-critical medical domain, with frequent decisions being made rapidly within a short period of time (on average 20 to 30 minutes) (Faraj & Xiao, 2006). This fast-paced and critical nature of work requires an overall situation awareness of the patient status and team activities during resuscitation. To maintain situation awareness, team members must obtain and share a large amount of diverse information in a timely fashion. Even so, information gathering and sharing are currently minimally supported by information and communication technology (ICT). Trauma teams primarily rely on verbal communication to share information and team leaders rely on short-term memory for information retention. Although introducing ICT in this domain poses many challenges—it may, for example, require changes in the current work practices—we believe that implementing ICT could facilitate information flow, improve situation awareness, and support decision making.

Our long-term research goal is to design and develop ICT solutions to improve situation awareness in high-risk and safety-critical medical settings. In this poster, we focus on designing a system to allow for acquisition and integration of heterogeneous information, and to present the integrated information to trauma teams in real time. The key contributions of this poster are: (1) system requirements for technologies to support situation awareness and decision making in an emergency medical domain. (2) Initial design of a prototype system for real-time information acquisition, integration and presentation.

Related Work

Information technology has been used for optimal data acquisition, integration and presentation in a number of medical and non-medical domains (Bardram et al., 2006; Heath & Luff, 1992; Hourizi & Johnson, 2001; Hutchins, 1995; Mumaw et al., 2000). Despite these efforts, trauma resuscitation remains one of the few information-intensive work environments with minimal ICT support. Several attempts have been made to introduce computer-based decision support systems in resuscitation areas, but they have not yet yielded optimal outcomes (Berlin et al., 2006; Fitzgerald et al., 2011). The key reasons for the lack of success include the challenges of capturing real-time data from diverse sources (Sarcevic et al., 2012), and integrating these disparate data in a dynamic environment (Barthell et al., 2004; Berlanga et al., 2008; Halevy, 2005). The increasing use of technologies, such as radio frequency identification (RFID)
(Fry & Lenert, 2005) and personal digital assistants (PDAs) (Anantharaman & Swee, 2001) in medical domains provides an opportunity for exploring automatic techniques for capturing data in real time. Large wall displays, on the other hand, allow for efficient data presentation and visualization, and the effectiveness of this approach has already been evaluated in several medical domains (Bardram et al., 2006; Bitterman, 2006; Parush et al., 2011). Coupling computerized data entry using medical devices with large wall displays offers a mechanism for providing additional teamwork support during trauma resuscitation.

System Requirements: Information Acquisition and Sharing

The initial design of our prototype system is based on in-depth field studies conducted at two Level I trauma centers over the past five years (Sarcevic & Burd, 2008; Sarcevic & Burd, 2009; Sarcevic et al., 2012). These studies revealed several challenges in information acquisition and sharing during resuscitations.

Incomplete and Missing Information

En route communication between Emergency Medical Services (EMS) paramedics transporting the patient and the emergency department (ED) receiving the patient is a critical source of information about the incoming patient. This pre-hospital information is used to formulate initial decisions about treatments, resources, and needed personnel (Sarcevic & Burd, 2009; Sarcevic et al., 2012). Our field studies have shown that only a small portion of this information reaches the trauma team leader before patient arrival; the rest is available from the EMS report upon patient arrival. The trauma leaders we interviewed expressed the need for more detailed and timely pre-hospital information to aid their initial decisions. This finding led us to consider approaches for automatic capture and display of pre-hospital information as it is being communicated en route.

Information Overload and High Mental Workload

During resuscitation, the team collects and shares a large amount of information about the patient status and team activities. We found that a team leader manages about 60 different information types in a typical event, including patient demographics, initial status, injury and medical history, vital signs, evaluation findings, and patient reactions to treatments (Sarcevic & Burd, 2008; Sarcevic et al., 2012). Interviews with team leaders also showed that managing multiple patients increases the difficulties in gathering and retaining patient information (Sarcevic et al., 2012). Because there are no mechanisms by which this information is acquired, integrated and stored, the leaders mainly rely on their working memory to process information. Sole reliance on working memory, however, increases the mental workload of team leaders, leading to diagnostic errors. This finding led us to consider mechanisms for providing external memory aids to support rapid acquisition and processing of information.

Inefficient Information Capture Techniques

The current low-tech techniques for capturing information during trauma resuscitation cannot meet the needs of the rapidly changing environment. For example, fluid dosage administered to the patient requires constant monitoring. In the current practice, however, the nurse responsible for administering medications also monitors the progression of the fluid drip, making the information about the fluid status difficult to obtain in a timely fashion. More recently, RFID technology has been adopted for monitoring medication and fluid administration in critical care units (Ohashi et al., 2008). We believe that high-tech approaches such as RFID have the potential to increase efficiency and accuracy of information capture in trauma resuscitation.

Based on our findings, we propose a computerized decision-support system that can acquire, integrate and present critical patient information in real time to address the identified challenges and support teamwork during trauma resuscitation.
A Prototype System for Managing Information during Trauma Resuscitation

The system architecture is the technological core that allows the optimal management of information (Hernando et al., 2008). The architectural design should make the system easy to deploy, scale, maintain, and enhance. The most commonly used application for developing system architecture is a component-based, three-tier model, which provides several benefits, including reusability, manageability, flexibility and scalability (Chu & Cesnik, 2000; Komatsoulis et al., 2008). This model divides a system into three subsystems or tiers: (1) the client tier provides a user with access to the system, (2) the data tier stores data in a relational database management system (RDBMS), and (3) the middle tier defines rules and algorithms for data processing.

Given the complexity of our application domain, we found the three-tier model the most appropriate for architectural design of our system (Figure 1). We selected a web-based application for the client tier because it allows portability and flexibility on the client side when accessing the system. For the middleware application, we adopted Extensible Markup Language (XML) following the Simple Object Access Protocol (SOAP) standard. By doing so, users can use a web client efficiently to communicate with the RDBMS.

The core component for processing and managing data in our system consists of a database, a Web server, and middleware applications located on the hospital server (Figure 2). This core system component is dedicated to processing incoming data from medical devices and providing the synthesized data to trauma teams via the display. The current system integrates three medical devices—EMS terminal, RFID scanner, and Lab computer—for obtaining different data types and transmitting them to the server (Figure 2). The integration of additional devices is planned for the future. In particular, the EMS terminal allows EMS paramedics to enter pre-hospital data that can be fed directly to the system and shown on a large wall display in the emergency room. RFID scanner collects data from the environment, including objects in use and administered medications and fluids, and presents this information directly to the team via the display. The lab computer is used for entering lab results, such as those obtained from blood draw, which are then fed directly to the system and shown on the display.

Conclusion and Future Work

The proposed system has the potential to support the work of trauma teams by acquiring, synthesizing and presenting critical patient information in a timely fashion. The system can also be used as an external memory aid to help trauma teams, and leaders in particular, acquire and process information needed for decision making.
Our next step is to assess the feasibility of the proposed system using a series of focus groups with trauma team members. In doing so, we plan to gather additional design requirements and to evaluate the extent to which the system affects current work practices. In addition, we will address several technical challenges, including (1) accurately extracting useful information from unstructured or semi-structured communication data, (2) semantically integrating heterogeneous data from different devices and sources, and (3) efficiently retrieving and presenting information queried by trauma team members.

**References**


Mumaw, R.J., Roth, E.M., Vicente, K.J., & Burns, C.M. (2000). There is more to monitoring a nuclear power plant than meets the eye. Human Factors 42, 1, 36-55.
Integrating Data Curation Concepts Throughout the Project Lifecycle: A WILIS Case Study

Cheryl A. Thompson  
Graduate School of Library and Information Science  
University of Illinois  
cathmps2@illinois.edu

Joanne Gard Marshall  
School of Information and Library Science  
University of North Carolina at Chapel Hill  
marshall@ils.unc.edu

Susan Rathbun-Grubb  
School of Library and Information Science  
University of South Carolina  
srathbun@mailbox.sc.edu

Jennifer Craft Morgan  
Gerontology Institute  
Georgia State University  
jmorgan39@gsu.edu

Amber Wells  
Sociology Department  
University of North Carolina at Chapel Hill  
awells82@email.unc.edu

Abstract

Researchers and funders continue to be concerned about the lack of archiving of scientific data. A web-based guide to data archiving for LIS researchers is being undertaken as part of the WILIS project. The WILIS project consists of: 1) an in-depth retrospective career survey of graduates of LIS programs in North Carolina with 2,653 respondents; 2) a modified recent graduates’ survey that was tested in 39 LIS programs in North America with 3,507 respondents; and 3) the preparation and archiving of the WILIS datasets in a publicly accessible data archive. This experience is being used to create the guide to data archiving. The guide will contain a full description of the steps involved in preparing and archiving datasets such as data cleaning, de-identification, preparation of supporting documentation, metadata, submitting to an archive and marketing the availability of data. The guide will be useful to researchers, data managers and data archivists/librarians.

Keywords: data curation, archives, LIS workforce

Introduction

There continues to be concern among researchers and funders about the lack of archiving of research data. Such data can be a useful for secondary analysis and for readers of published research articles who may be interested in further information or replication studies. Most research data sets are not archived at all and, when data archiving is considered, it is often an afterthought, which makes successful data archiving difficult, if not impossible.

This poster will discuss the curation of large-scale survey data collected as part of the Workforce Issues in Library and Information Science (WILIS) program of research funded by the Institute of Museum of Library Services (IMLS) from 2005 to 2013. The WILIS projects were designed to study the educational, work, career and retention issues faced by library and information science (LIS) graduates. WILIS has been a partnership of the University of North Carolina at Chapel Hill (UNC) School of Information and Library Science and the UNC Institute on Aging.

Acknowledgements: The WILIS 1, 2 and 3 studies were supported by grants from the Institute of Museum and Library Services. The primary research team for the WILIS3 study at the University of North Carolina at Chapel Hill includes: Joanne Gard Marshall, School of Information and Library Science, Lead Principal Investigator; Jennifer Craft Morgan, Co-Principal Investigator, UNC Institute on Aging; Jonathan Crabtree, Co-Investigator, Odum Institute for Research in Social Science; Susan Rathbun-Grubb, Co-Investigator, University of South Carolina; Cheryl A. Thompson, Project Manager; Marshica Stanley, Social Research Assistant; Brian D. Leaf, Graduate Research Assistant; Thu-Mai Christian, Graduate Research Assistant; and Amber Wells, Doctoral Research Assistant.


Copyright is held by the author/owner(s).
Phase 1 of WILIS consisted of an in-depth, retrospective study of graduates of LIS programs in North Carolina from 1964-2007. Phase 2 modified the recent graduates’ portion of the WILIS1 survey and tested this survey tool for shared alumni tracking in 39 LIS programs in North America. Phase 3 of WILIS is currently focusing on archiving the WILIS1 and WILIS2 data for future use by researchers and other stakeholders and creating a guide to data archiving that can be used by LIS and social science researchers. Lessons learned from the WILIS data archiving experience form a useful case study for developing best practices for data curation throughout a research project.

The WILIS Case Study

The results of the WILIS projects have been reported as part of two theme issues of Library Trends in Fall 2009 and Fall 2010. The in-depth, retrospective survey of LIS graduates in North Carolina conducted between 2005 and 2008 were reported by Marshall et al (2009). The results of the shared alumni tracking student in 39 LIS master’s programs was reported in Library Trends Fall 2010 (Marshall et al, 2010).

The goals of the WILIS 3 project (2010 - 2013) are to: 1) create publicly accessible de-identified datasets from the WILIS studies; 2) develop an interactive system that will enable users to explore the WILIS data; and 3) produce a best practices toolkit for data archiving that can be used by other researchers. The study data will be deposited at the Odum Institute for Research in Social Science at UNC, a member of the Data Preservation Alliance for Social Sciences (Data-PASS). Federated data archives, such as Data-PASS, provide secure depositories that are networked to provide a back-up storage system, so as not to rely on a single server, and to reduce the risk of preservation loss.

The Odum Institute preserves and provides access to data via the Dataverse Network developed by Harvard University. More detail about the Dataverse is available at http://thedata.org/. The Dataverse technology offers researchers a solution for publishing their research data and providing access to users. It allows researchers to preserve data files and study documentation in any file format; if data is in Stata or SPSS format, it will generate variable metadata automatically. The metadata template in the Dataverse is compliant with the Data Documentation Initiative (DDI) but metadata can be exported into other formats such as Dublin Core, FGDC and MARC. While the indexing system enables searching by allowing multiple access points and harvesting, researchers have control over the access and use of their study data. The Dataverse allows researchers to subset and analyze data within the web interface as well as download files in multiple formats. This archiving technology also was appealing to the WILIS team because it generates a formal citation for each data file, including a persistent identifier and URL, allowing researchers to cite and receive credit for their work.

Conclusion

Traditionally, the research data lifecycle has involved data planning, production, management, analysis and storage; however, new models for curating data across the lifecycle have emerged (DCC, 2011; Choudury, 2010). Researchers need practical advice in order to apply these new curation models to their project and data workflows. WILIS 3 has documented the process of planning and archiving the WILIS data and offers a model for how researchers can prepare for archiving earlier in the research process. For example, the WILIS 3 Guide (in press, 2013) identifies key considerations at different points in the project life cycle and walks the reader through the process of identifying appropriate documentation to be included in the data archive to maximize future usefulness of the data. The WILIS archiving case study illustrates the type of information that needs to be documented as the researcher moves through each stage of the research process. The poster will discuss lessons learned from our experience such as the importance of documenting methodological decisions throughout the research process, understanding the implications of data structure (e.g. open response text) for future de-identification, developing a data management plan and selecting an archive or repository.

The WILIS 3 Guide approaches data archiving from the point of view of three stakeholders: the researcher, the data manager and the data archivist/data librarian. An interactive, web-based user-interface and documentation scheme for the WILIS study data will be used to illustrate how one might provide enhanced context for secondary users of library and information science and other social science data. The WILIS 3 Guide provides a model for integrating archiving throughout the research process and highlights best practices for designing and implementing a data management plan throughout the data
lifecycle. Ultimately, better planning for archiving may result in sharing higher quality data packages. The resulting package will also make the data more easy to use for researchers and other stakeholders, thus maximizing the funder's return on investment and secondary use of the data.

References


Interacting with Health information for Self-care: A Pilot Study Exploring Undergraduates' Health Information Literacy

Jinxuan Ma  
Florida State University  
jm07@my.fsu.edu

Don Latham  
Florida State University  
dlatham@fsu.edu

Abstract

This poster presents findings from a dissertation pilot study of how undergraduate students apply their knowledge and skills to find, evaluate, and use health information for self-care. The study employed Dervin’s sense-making theory and examined to what extent undergraduate students demonstrate health information literacy (HIL) knowledge and skills within the context of seeking and using health information for self-care. It utilized a three-phase design: (a) a survey, (b) an interview, and (c) a follow-up observational study. Based on purposive sampling, 34 undergraduates at Florida State University participated in the survey, and eight of them were selected as participants for the interview and the observational study. The study results showed major barriers students encountered and strategies they used in various situations when seeking and evaluating information for self-care. This study has implications for designing more effective college health promotion programs and developing objective HIL assessment instruments in the context of health information seeking and use.

Keywords: consumer health information seeking behaviors, health information literacy, self-care, Dervin's sense-making theory

Introduction

College life is a transitional time for undergraduate students for taking charge of their own health and maintaining a healthy lifestyle, which is essential to their academic achievement and their overall adult health outcomes. Existing studies show that college students, especially undergraduate students, may not have sufficient health literacy skills (Ickes & Cottrell, 2010; Ivanitskaya, O’Boyle, Casey, 2006). This exploratory study employed Dervin’s sense-making theory (1983, 1992) to explore how undergraduate students apply their HIL knowledge and skills to find, evaluate, and use health information for self-care. It also implemented a three-phase study design for data collection: (a) a survey, (b) an interview, and (c) a follow-up observational study. The study results are expected to contribute both theoretical and practical knowledge to the current limited body of research on undergraduate students' health information-seeking behavior and health information literacy.

Background

Health information literacy refers to the ability to recognize health information needs and to apply the knowledge and skills required to find, evaluate, and use information to make health-related decisions (MLA, 2003). Health information literacy knowledge and skills range from basic comprehension of evidence-based medicine, to knowledge of credible information sources, to core information skills supporting informed decision-making. Such knowledge and skills are crucial for health consumers to efficiently seek and use health information for self-care.

Acknowledgement: This work was sponsored by a Dissertation Research Grant for Fall 2012, from The Graduate School, Florida State University.


Copyright is held by the authors.
Previous studies have stressed the importance of health information literacy and the role of information professionals in promoting HIL (Banas, 2008; Ivanitskaya, O’Boyle, Casey, 2006; Putnam, Kitts, Pulcher, 2010; Schardt, 2011; Shipman, Kurtz-Rossi, & Funk, 2009). However, a lack of HIL assessment instruments is problematic to existing HIL studies. Even though HIL is recognized as a key component of information users’ engagement with health information for health purposes (Yates et al., 2012), little research has explored to what extent those users demonstrate their HIL knowledge and skills in identifying their information needs and in seeking and using health information for self-care. There is also limited understanding about how HIL knowledge and skills affect the ways in which individual users seek and use health information in specific situations.

**Methods**

This study asks one overarching research question with four sub-questions: To what extent do undergraduate students demonstrate HIL knowledge and skills in their health information seeking and use for self-care? (a) What are the situations that cause students to seek health information for self-care? (b) What gaps do students encounter in seeking health information for self-care? (c) How do undergraduate students bridge the gaps in seeking health information for self-care? (d) How do undergraduate students use health information that they have obtained for self-care?

To answer the research questions, this study employed a three-phase design: (a) a group-administered survey, (b) a semi-structured interview, and (c) an observational study of online health information searching using the think-aloud protocol method.

The target population was undergraduate students enrolled at Florida State University who are actively seeking and using health information for self-care. The group-administered survey recruited 34 participants from two undergraduate courses. The survey collected data about (a) demographic information, (b) basic health status, and (c) health information seeking behaviors. Eight participants for the interview and observational study were selected from the survey participants, based on their willingness to participate and their interest in seeking and using health information for self-care. The interview questions were informed by Dervin’s sense-making theory to explore individual information users’ experiences in their health information-seeking and use for self-care. The follow-up observational study was designed to complement the interview data by examining students’ online health information search skills for answer six self-care scenario questions. The questions covered a variety of undergraduate students’ self-care questions including (a) major recommended vaccines for freshmen, (b) on-campus sources for treating eating disorders, (c) available helpful information for stress and sleeping disorders, (d) coping strategies for dealing with side effects of antidepressants, (e) potential harm associated with binge drinking, and (f) local HIV/AIDS testing services (Questions c &d were adapted from Buhi et al., 2009). The questions were pretested by several health professionals in the student health center at Florida State University. A think-aloud protocol was used to examine how participants seek and use online health information for self-care. Camtasia Studio 7.0 was used to capture participants’ thoughts and search processes while answering the six questions. After the data collection, NVivo qualitative data analysis software was used to analyze data from both the interview and the observational study.

**Findings**

To prevent disease or control minor health problems, the participants largely relied on seeking health information for self-care (e.g., symptoms, side effects of medicine). Three major sources were mentioned by the participants: Internet (e.g., commercial, organizational, educational, and government websites), people (e.g., parents, friends, and health professionals), and print materials (e.g., pamphlets and magazines). Google was commonly used as a starting point for online health information seeking, even for searching their university health services. They rarely understood the limitations associated with their frequently used Google search strategies—basic keyword search and question inquiry—even though most participants stated they did not have problems in searching for online health information. Exchanging information with friends and consulting friends (especially those in health-related majors) about nutrition, diet, health problems, or medications were preferred strategies of seeking health
information. Before seeking help from health professionals, they tended to consult their parents. Source accessibility and affordability were two major barriers concerns for source selection and use.

Many criteria (e.g., relevance, reliability, accuracy, trustworthiness, credibility) for evaluating health information were identified from the data despite students' different understanding of these criteria. They frequently selected the first Google search result or the top five results as satisfactory answers without any further scrutinizing. Additional ways of judging information were common sense, familiarity, personal experience, multiple sources, and consumer reviews.

The participants articulated what barriers to health information seeking and use they encountered and how they bridged these barriers. Two major barriers were their insufficient health literacy skills including terminology and background knowledge, and weak information searching skills, which included difficulties with search term selection and poor skills in searching databases, electronic journals, and other online health information resources. Some participants also found it difficult to determine source authority, accuracy, and reliability due to information overload and source disagreement. They tended to use different ways of bridging these barriers, such as verifying information through personal experience, multiple sources, common sense, or consumer reviews. Consulting people (e.g., parents, friends, health professionals) and the Internet were mentioned as complementary strategies. They may also give up on their efforts to judge information because of frustration and lack of time.

The participants tended to use health information to make health-related decisions, such as changing behaviors, undertaking self-diagnosis, engaging in self-treatment, or seeking professional help. They also applied the information they obtained to prevent diseases by practicing good nutrition and diet management and to control health problems by monitoring existing symptoms or side effects of medications. When using health information for self-care, both positive and negative outcomes were reported in terms of controlling health problems, improving health conditions, making health decisions and clarifying questions.

The data of the observational study showed that most participants confused (a) consumer health information and professional health information, (b) reliable health information and information with uncertain quality, and (c) digital fluency and health information literacy.

Conclusion

Many of the findings verified the results of prior research about undergraduate students' insufficient health information literacy knowledge and skills within the context of seeking and using health information for self-care. However, this study provided more evidence about major barriers that prevented students from seeking and using reliable information for self-care, strategies they used in specific self-care situations, and outcomes of their information use. It is valuable to integrate such understanding in designing more effective college health promotion programs and developing objective HIL assessment instruments in the context of health information seeking and use.

References


A Comparative Study of Data Reuse Among Quantitative Social Scientists and Archaeologists

Ixchel M. Faniel  
OCLC Research  
fanieli@oclc.org

Julianna Barrera-Gomez  
OCLC Research  
barreraj@oclc.org

Adam Kriesberg  
University of Michigan School of Information  
akriesbe@umich.edu

Elizabeth Yakel  
University of Michigan School of Information  
yakel@umich.edu

Abstract

This paper presents preliminary findings from a comparative study of data reuse in the quantitative social science and archaeology disciplines. Quantitative social scientists have been engaged in large-scale data sharing and reuse from centralized repositories for over 50 years. In contrast, archaeologists are transitioning from sharing and reusing data on a small-scale with colleagues and museums to large scale sharing and reuse via centralized repositories. In this study, we consider whether approaches to supporting data reuse in quantitative social science can be applied to the archaeological community. Currently we are examining data reuse practices in both disciplines via three points of comparison and will discuss preliminary findings regarding: 1) the nature of context needed during reuse, 2) the use of a bibliography of data related literature, and 3) the role of intermediaries.

Keywords: data reuse, data repositories

Introduction

A common data infrastructure, albeit necessary to support the central goals of eScience, may not arise given minimal research to discover common ground among different disciplines (Borgman, 2007). Comparative studies among eScience initiatives have been informative, but attention has tended to focus on degrees of disciplinary difference. We have learned how different levels of interdependence among scientists and task uncertainty have implications for design and use of eScience infrastructures and how the different histories and configurations of disciplines might impact how researchers contextualize and document research data and processes (Birnholtz & Bietz 2003; Carlson & Anderson 2007; Fry, 2006). However, we know less about the similarities that exist among disciplines and how they can be drawn upon to develop common approaches to support data reuse.

In this study, we explore data reuse in two disciplinary communities – quantitative social science and archaeology. Quantitative social scientists have been engaged in large-scale data sharing and reuse from centralized repositories, some of which have implemented procedures and standards and built expertise and reputations over 50 years (e.g. Inter-university Consortium of Political and Social Research). In contrast, the archaeological community is transitioning from sharing and reusing data on a small scale with colleagues and museums to large scale sharing and reuse via centralized repositories. They too have implemented procedures and standards, but also look to the well-established data repositories to inform their work (e.g. Open Context). A major objective of this study is to examine the data reuse practices in both disciplinary communities to determine whether approaches that support data reuse among the quantitative social scientists might be applied to the archaeological community. In the following paragraphs, we draw from the data reuse literature to discuss the three ways we are comparing data reuse practices. We then describe our research methods and end with a discussion of the preliminary findings and implications.

Acknowledgement: This research was made possible by a National Leadership Grant from the Institute for Museum and Library Services, LG-06-10-0140-10, “Dissemination Information Packages for Information Reuse”.  

Copyright is held by the authors.
Data Reuse Literature

This current study examines three points of comparison found in data reuse studies that have implications for the design, development, and use of data repositories. The first is what information about the context of data production is needed during reuse. Although studies agree that context is critical for reuse, there is disagreement as to whether enough can be captured to discover and evaluate reusable data (e.g., Faniel & Jacobsen, 2010; Zimmerman, 2008). The latter hinges on the nature of the context, particularly how it is generated and whether it is captured during data production, rather than on a specific discipline or data type (e.g. Carlson & Anderson, 2007; Jirotka et al., 2005). In separate studies, both field anthropologists and HIV/AIDS laboratory researchers were found to rely on “craft like methods” and tacit knowledge that were difficult to explicate through systematic documentation of the research process (Birnholtz & Bietz, 2003; Carlson & Anderson, 2007). The second point of comparison is the use of bibliographies of data related literature. As an additional means of documenting data, they serve multiple purposes during the data reuse process, including discovering relevant data, constructing systematic data samples, understanding how data have been conceptualized and measured, and keeping informed about community debates around the data (Faniel & Jacobsen, 2010; Faniel, Kriesberg, & Yakel, 2012; Zimmerman, 2007). Lastly, we compare the role intermediaries play between data producers and reusers. While a repository itself and the people who staff it serve as intermediaries, research has also found data reusers rely on other third parties, such as advisors and colleagues, to facilitate data reuse (Faniel et al., 2012; Jirotka et al., 2005).

Methods

Our findings are drawn from data collected during three rounds of interviews conducted between June 2011 and April 2012. In total, we spoke with 66 participants: 44 quantitative social scientists (22 novices and 22 experts) and 22 archaeologists. In our series of semi-structured hour-long interviews, we asked respondents to discuss their experiences reusing data in their particular field of research. Topics of inquiry included how respondents discovered and evaluated data for reuse and their experiences and thoughts about digital data repositories. All interviews were audio recorded and transcribed. We then coded the transcripts using NVivo, a qualitative data analysis software tool. To achieve inter-rater reliability, two members of our project team coded each group of transcripts.

Preliminary Results

The quantitative social scientists reused a broad range of survey data collected from human subjects and local, state, and federal institutions. They also reused quantitative data that had been transformed from qualitative data, such as measures of the “democratic-ness” of a country produced by analyzing newspaper articles. We found much of the context created for these data was digital (e.g. Microsoft Word documents, online surveys) and static. For instance, a typical research design was developed at the beginning of a project, not much was changed for the duration of the study, and data were collected in standard forms. In contrast, the archaeological data were more varied, consisting of images of objects found during a site survey or excavation, textual descriptions of those objects, geographic location data documenting objects within a site, or GIS shape files documenting a site through time. The context was more dynamic, evolving as the project evolved. During excavation of a physical site, documentation was generated daily and research plans were subject to change based on internal and external factors beyond the excavator’s control (e.g., the nature of excavation discoveries, weather conditions).

Despite these differences in the nature of the context, we found similarities between quantitative social scientists and archaeologists. Both groups expressed a need to understand the data producer’s research methods, especially aspects of context that provided insight into how the data producer carried out the research. Both acknowledged that data producers conducted research in different ways. Knowing more about research procedures helped data reusers understand the nuances and make more informed reuse decisions. For instance, quantitative social scientists wanted to know how data producers defined and measured the variables data were intended to capture. Archaeologists were interested in the type of GPS device data producers used to identify site and object location, since these instruments provide location information at different levels of specificity.
Quantitative social scientists also discussed bibliographies of data-related literature. They used articles written by data producers to get more detailed and clear descriptions about measurement and methods or justifications for research decisions. They also used articles written by community members to see how the data were critiqued and reused. Archaeologists also discussed bibliographies as both a source of data discovery, since raw data are often published in appendices, and as a gateway to locating related project data. They also referenced journal articles as a primary step in discovering both projects and published datasets that might be relevant for reuse. Additionally, for archaeologists bibliographies played a central role in locating associated project reports specialists wrote that included faunal analyses or ceramic typologies.

Quantitative social scientists and archaeologists discussed the role of intermediaries in discovering, evaluating, and understanding data. Social scientists cited the help of colleague networks and mentors in finding relevant data and understanding data limitations. Archaeologists sought advice from colleagues in locating relevant data for reuse, but also asked museum curators questions about the data to get a better understanding of its context when the data producers were not available.

Conclusion

This paper presents preliminary findings from a comparative study of quantitative social scientists and archaeologists. Our immediate plans are to continue data analysis to more fully consider study implications. However, early indications suggest that despite differences in how context was generated and captured during data production there were similarities as well differences in the reuse practices of quantitative social scientists and archaeologists. For instance, data reusers from both disciplines wanted context about the data producer’s research methods and were able to get enough detail to be able to reuse the data. Both also used bibliographies of data related literature, but for different proposes. Archaeologists relied on bibliographies to facilitate data discovery, whereas quantitative social scientists used bibliographies to facilitate reuse decisions. Data reusers in both disciplines also relied on intermediaries. However, different types of people were used for different purposes. Quantitative social scientists, particularly novices, relied on faculty advisors who had more data reuse experience to find relevant data and understand data’s limitations, whereas archaeologists relied on colleagues and museum curators to locate data and associated context.

References


Authenticity as a Social Contract—We Are Our Records

Corinne Rogers  
University of British Columbia  
cmrogers@mail.ubc.ca

Joe Tennis  
University of Washington  
jtennis@uw.edu

Abstract

This poster presents the authors’ work to date on developing an application profile for authenticity metadata (the IPAM, or InterPARES Authenticity Metadata), including (1) the functional requirements, (2) metadata elements derived from the Chain of Preservation model from the InterPARES research project, (3) a crosswalk of a sample of IPAM elements to Dublin Core, PREMIS, and MoReq2010, (4) those elements deemed essential to presume the authenticity of a record as it moves from creation to permanent preservation, and (5) next steps, integrating the application profile into the Archivematica preservation system the core elements of the application profile relating to maintaining the presumption of authenticity through preservation and access.

Keywords: metadata, authenticity, social contract, application profile

Records are defined in archival science as documents made or received in the course of practical activity and set aside for further action or reference (Duranti, 2009). Records are shared across space and over time, participating in complex social interactions. Their provenance and relationship with the actions and functions in which they participate or which they document involves them in a social contract. Social interaction, whether in person or online, involves a social contract between participants that is based on trust.

It is commonly accepted that the records that document our actions and transactions, and the data that fuels the information highway must be trustworthy. Trustworthiness depends on our ability to assess many elements, including authenticity of the record (that is its identity and integrity), its reliability (a measure of its completeness and the controls exercised on the process of its creation), and its accuracy (the truthfulness, precision, and completeness of its data content) (Duranti, 2009). Maintaining the presumption of authenticity of digital records is a pressing concern for society.

Digital records are the lasting traces of actions and transactions, and their evidentiary capacity depends on their authenticity, reliability and accuracy, established at creation, and maintained and preserved over time and across technological change. Metadata, the machine- and human-readable assertions about information resources, enables intellectual, physical, and technical control over information resources, that is, it enables the trustworthiness of records. It is primarily discussed in the context of information retrieval and preservation. However, metadata provides key elements establishing record authenticity – that is, identity and integrity. While functional requirements for preservation metadata are transitioning from research to standardization and implementation (e.g., PREMIS), little work has connected archival theory with functional requirements for authenticity metadata (cf. MoReq, ISO, AGrKMS). However, there is a wealth of research into the requirements for authenticity and reliability of records at their creation, and throughout their life cycle, culminating in preservation (Duranti & Preston, 2008). Many extant metadata schemas account for identity and integrity (the components of authenticity), but do not explicitly capture the attributes identified and required by the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) project. And although we may be able to repurpose some aspects of metadata schemas to capture identity and integrity metadata, it is incomplete (according to InterPARES requirements and by extension, diplomatic analysis), and risks being confused with the original purpose of the repurposed metadata schema. This would violate the requirement for interoperability established within the context of InterPARES 2 and codified in InterPARES 3 (Tennis, 2010; Tennis & Rogers, 2012).

Rogers, C., & Tennis, J. (2013). Authenticity as a social contract—We are our records. iConference 2013 Proceedings (pp. 801-803). doi:10.9776/13393
Copyright is held by the authors.
We achieve confidence in records – the products, byproducts, and residue of our activities and transactions online – through a balance of mechanisms of trust and control (Cofta, 2007). Trust in records is based on what we know about their creator and/or their custodian: their reputation, evaluated on past actions and conduct; their performance, or the relationship between the trustee’s present actions and the conduct required to fulfill his or her current responsibilities as specified by the trustee; competence, which consists of having the knowledge, skills, talents, and traits required to be able to perform a task to any given standard; and confidence, which is an “assurance of expectation” of action and conduct the trustee has in the trustee (Borland, 2009; Duranti & Rogers, 2011; Sztompka, 1999).

The determination and assessment of authenticity depends on the circumstances of record creation, and framework of subsequent preservation. According to InterPARES, to assess the authenticity of a digital object, one must be able to establish its identity and demonstrate its integrity. The identity of a digital object is established by the attributes of the object that uniquely distinguish it from other objects, while integrity refers to its wholeness and soundness, that is, the degree to which it is complete and uncorrupted. InterPARES recognized and articulated the difference between the form in which a document is viewed by a person reading it, and that in which it is stored in the electronic system. The layers of abstraction introduced by the technology between the physical and logical record have implications for the assessment of authenticity. This significant difference between paper and digital records is at the root of the challenges of integrating digital records with the legal system’s conception of documentary evidence.

This poster visualizes the authors’ work to date on developing an application profile for authenticity metadata (the IPAM, or InterPARES Authenticity Metadata). This is design research. As design research, the methodology followed a course of knowledge acquisition: (1) establish the desired outcome for metadata schema; (2) follow the DCMI Singapore Framework; (3) develop functional requirements and model those through entity relationship diagrams in order to then construct an IP3 application profile; and (4) vet this with stakeholders, researchers, and the DCMI community. The research was undertaken through the lens of archival theory in general, and the findings of InterPARES 1 and 2 specifically, applied in relation to the Guidelines for Application Profiles and the Singapore Framework, both issued by the DCMI. We were guided by literature on the creation of application profiles (Heery & Patel, 2000), and examples of application profiles (Collections and Scholarly Works Application Profiles).

In this poster we present five things. (1) We begin with the functional requirements of authenticity metadata and entity relationship diagrams of the modeled entities. (2) We then offer a graphical representation of the metadata elements derived from the Chain of Preservation model from the InterPARES research project. (3) These elements are then crosswalked to an illustrative sample of IPAM elements to Dublin Core, PREMIS, and MoReq2010. (4) Then we highlight those elements deemed essential to presume the authenticity of a record as it moves from creation to permanent preservation. We organize the material in the form of a visualization of the records life cycle, showing insertion points for metadata at key moments of creation, capture, appraisal, disposition, and preservation. Finally, (5) we describe upcoming and continuing work in collaboration with Artefactual Systems (http://artefactual.com/). This work integrates the application profile into the Archivematica preservation system using elements relating to maintaining the presumption of authenticity through preservation and access.

References


Randomized Sampling:
An Approach to Extraction of Metadata Records

Olajumoke Azogu  
University of North Texas  
olajumoke.azogu@unt.edu  

Jiangping Chen  
University of North Texas  
jiangping.chen@unt.edu

Abstract

Random samples are desired by researchers to produce results that are generalizable. We present an approach to extract random samples of metadata records from Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) repositories and others such as Z39.50 collections. Based on our approach, we developed PHP scripts to randomly extract 2000 metadata records from two digital repositories. Our approach can be used by others that need to extract representative samples from digital collections.

Keywords: algorithm, metadata, harvesting, extraction, oai-pmh, Z39.50

Introduction

Random sampling remains one of the preferred sampling methods for scientific research because it can generate representative samples of populations of interest, therefore guarantees the generalizability of results. With ever increasing digital objects and growing interests in investigating better digital content management, extracting representative samples through random sampling is important and efficient for working with large collections.

We present our approach to randomly extract any number of metadata records from digital collections. This approach was used to extract random samples for the MRT Project\(^1\) from the Catalog of the University of North Texas Libraries (UNTL) and the Portal to Texas History (PTH) digital library. These two digital collections use different metadata standards.

Metadata Harvesting and Extraction Approaches

Our approach stemmed from existing methods for metadata harvesting; the process by which digital collections’ metadata records can be accessed and extracted through established interchange standards among collections that enable sharing (Arms, Dushay, Fulker & Lagoze, 2003). We explored extraction across the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Z39.50 protocol.

OAI-PMH-based Harvesting/Extraction

The OAI-PMH developed for interoperability between digital collections (http://www.openarchives.org/pmh/), has been broadly acknowledged as a means for metadata harvesting (Van de Sompel, Nelson, Lagoze & Warner, 2004). It supports the identification and extraction of metadata records, and dissemination of records in various metadata formats. However, to ensure

---

\(^1\) The title of this project is “Enabling Multilingual Information Access to Digital Collections: An Investigation of Metadata Records Translation,” a research project sponsored by the Institute of Museum and Library Services (IMLS) and the University of North Texas (UNT) that aims to evaluate the extent to which current machine translation technologies generate adequate translation for metadata records, and to identify the most effective metadata records translation strategies for digital collections.

Acknowledgements: Institute of Museum and Library Services (IMLS) Grant LG-06-10-0162-10. We thank UNT Libraries for allowing us to use their digital collections. We also thank Mark Phillips and Jason Thomale for their help on this work.


Copyright is held by the authors.
interoperability, unqualified Dublin Core is required and outputs from databases must be in XML (Warner, 2001).

Several approaches and tools exist that use OAI-PMH to extract metadata from a digital collection. Typical examples like MarcEdit (http://people.oregonstate.edu/~reeset/marcedit/html/index.php) and ZMARCO (http://zmarco.sourceforge.net/) either harvests all records, specified set of records, or an individual record in a collection (http://www.openarchives.org/pmh/tools/tools.php). Additional solutions for metadata extraction include approaches where all records from a digital collection are harvested and then systematic sampling applied to draw representative samples. However, these were found to be ineffective for the purposes of truly randomized metadata records extraction or inefficient when dealing with huge digital collections that contain millions of records, when only relatively few, representative number of records were needed.

**Z39.50-based Harvesting/Extraction**

Z39.50 was developed for the search and retrieval of information in databases (ANSI/NISO, 2003). A rapid assessment of Z39.50 tools (http://www.loc.gov/z3950/agency/ resources/software.html) reinforced the need for a flexible, adaptable application similar to that to be developed for the OAI-PMH-compliant catalog. Not unexpectedly, available Z39.50 applications, in a similar way to OAI-PMH tools, did not fulfill the needs of the MRT project.

The MRT project needed a solution that could effectively and efficiently extract specified numbers of records as a representative sample of a digital collection. In addition, it required the isolation of six (6) elements per metadata record. Existing OAI-PMH or Z39.50 tools could not achieve this. Our solution therefore contributes an original approach to the harvesting of metadata records, which provides users with control and flexibility over the number of records and metadata elements needed from digital collections. In this approach, we randomize extraction to ensure that representative metadata records are obtained irrespective of the number of records to be extracted from a collection.

**Randomized Extraction of Sample Metadata Records**

For the purposes of the MRT project, 2000 metadata records were to be extracted from the PTH and UNTL Catalogs. The extraction processed metadata records with two different standards: DC and MARC. As a result, six elements (Title, Publisher, Description, Subject & Keywords, Coverage and Creator) per metadata record were extracted.

**The Approach**

Our approach follows these steps, starting with the base URL of the collection;

1. Obtain collections’ metadata records identifier header and ID number range;
2. Generate randomized IDs that fall within collection’s range. Repeats if ID has been generated, else continues;
3. Extract record with generated random ID;
4. For the MRT project’s unique need, six metadata elements of interest are isolated and saved;
5. Repeat steps (2), (3) and (4) until desired numbers of records are extracted.

Our scripting language was chosen for its functionalities to interface with both OAI-PMH and Z39.50 protocol servers. The above algorithm is illustrated in Figure 1.
Extraction from the Portal to Texas History (PTH) Digital Collection

The OAI-PMH-compliant PTH collection, contained about 133,000 metadata records (http://texashistory.unt.edu/, January, 2011). Using OAI-PMH verbs, and PHP, we executed our algorithm to extract 1000 random metadata record samples.

Two PHP scripts were written for this process. While the repository’s URL is needed and serves as entry point for the extraction process, the first script uses OAI-PMH verb ListIdentifiers to retrieve the repository’s ID header, and lower / upper boundaries of the metadata records IDs. The information is passed on to the randomized extraction process as parameters. The second script uses the same parameters and the repository’s URL to extract the number of records needed. It generates a random unique ID with the OAI-PMH verb GetRecord to access the corresponding metadata record. The extracted record is validated, six metadata elements obtained, multiple metadata element occurrences concatenated, and then written into database tables. Our script uses PHP’s mt_rand() function (http://php.net/manual/en/function.mt-rand.php) to generate random numbers and validates metadata records by checking for non-empty records and for those containing at least four of the six elements of interest.

Extraction from the UNT Libraries Catalog

We adapt our algorithm to extract 1000 records from the 1.9 million records (January, 2011) Z39.50 UNTL Catalog. A script, using PHP and its YAZ extension (YAZPHP), obtains the repository URL, and runs similar to the OAI-PMH extraction except that YAZPHP functions (yaz_connect, yaz_search and yaz_record) are used for extracting records. In contrast also, upper and lower record ID limits were manually obtained and hard coded into the script.
The PHP scripts can be found at [http://txcdk-v10.unt.edu/MRT/Scripts/welcome.html](http://txcdk-v10.unt.edu/MRT/Scripts/welcome.html).

**Experiments and Results**

To affirm the validity and randomness of the extraction process, two experimental runs were carried out, and a comparative analysis of the results carried out. Runs were conducted over a Linux server via a command line interface, and outputs stored to MySQL tables and plain text files. The two sample batches of 1000 each, more than statistically significant sample sizes, were determined not to be significantly different, making us conclude that our random samples are a sufficient representative of the entire catalog metadata records.

Test runs on three OAI-PMH repositories and three Z39.50 repositories were successful. However, several OAI-PMH ([http://gita.grainger.uiuc.edu/registry/ListAllRepos.asp](http://gita.grainger.uiuc.edu/registry/ListAllRepos.asp)) and Z39.50 repositories ([http://www.loc.gov/z3950/](http://www.loc.gov/z3950/)) were inaccessible either because of broken links, or restrictions that require prior, written approval for access. Access is therefore important for successful metadata records extraction.

**Discussion and Conclusion**

We developed a generalizable solution for randomized metadata records extraction that is freely accessible. Other advantages include flexibility and control over the number of metadata records and elements extracted, and extensibility of the script to suit unique uses. While our scripts can be run “as-is”, repository-specific parameters would necessitate changes. However, our solution makes these necessary changes easy.

Despite our accomplishment, work still needs to be done to extend the generalizability of our solution. The command-line interface could be replaced with a user-friendly graphical user interface (GUI) which would negate the need for server-end scripting. Our work stemmed from the need for flexibility in extracting metadata records. Even greater flexibility may be achieved by merging our solution with other harvesting solutions, for such needs as; extracting random records over time periods for quality assurance, or, domain specific extractions which might be useful to research interests. Possibilities are numerous, and we conclude that adaptable metadata records extraction is unfinished.

**References**


Organizational Perspectives of Open Innovation in Government

Claudia A. Louis
Syracuse University
calouis@syr.edu

Abstract

This study seeks to explore the factors and conditions that influence the adoption and implementation of open innovation practices in public sector agencies. It examines the implementation of a web-based crowdsourcing platform used to solicit public input in solving problems encountered by agencies. Theoretical development is based on open innovation literature in private firms and draws on classic innovation literature from both the private and public sectors. Qualitative data is collected through interviews and analyzed using a combination of deductive and inductive coding. The expected outcome is the deeper understanding of the influencing factors and conditions, as well as the various characteristics of these practices.

Keywords: open innovation, crowdsourcing, public sector innovation, contests

Introduction

The promotion of openness and transparency in government has been an area of focus for the Obama administration, as detailed in the President’s memorandum upon taking up office in January 2009 (White House, 2009). As part of the Open Government Initiative, the web-based crowdsourcing platform (challenge.gov) was developed to facilitate the solicitation of citizen input in solving problems encountered by the various agencies. Problems are formulated as competitions (referred to as contests and challenges) and offer rewards to the winning submission(s).

As an emerging practice it is important to understand the characteristics of the various contests and challenges, how they fit into the existing organizational structure of hosting agencies, and what environment and conditions promote their use. The specific research questions are: i) How is the crowdsourcing platform challenge.gov being used by the various federal agencies? ii) What factors influence the decisions to use this platform within the various agencies?

Conceptual Development

Openness in innovation refers to the practice of embracing outside participation as well as sharing with others on the outside (Chesbrough, 2011). Organizational boundaries are opened up to allow knowledge in-flows and out-flows, and create an atmosphere for combined innovation activities involving partners, customers, suppliers (Enkel et al., 2009). The opening up of organizational boundaries is enabled by the internet which allows the easy flow of information and the connection of widely dispersed experts and enthusiasts (Chesbrough, 2003).

Crowdsourcing refers to the practice of outsourcing to crowds. As an alternative to outsourcing to specific suppliers or consultants, firms make the details of their problems publicly available, and invite potential solvers to submit solutions. Solutions are usually solicited using a ‘broadcast search’, and are open to anyone who feels they are qualified to do so (Jeppesen and Lakhani, 2010). The crowdsourcing application being investigated can be classified as an instance of open innovation where ideas from outside the organization are used to solve internal problems.
Open innovation in the public sector represents an emerging area of research where distinguishing theories are yet emerge. This study combines literature from three main areas: open innovation from the private sector; and classical innovation from both private and public sectors. It proposes that factors influencing the decision to implement open innovation practices in public sector agencies operate at three distinct levels: organizational, project, and individual. The proposed factors have been selected due to their prominence in the literature and are examined below:

Organization Level:
- Bureaucracy and red-tape: degree of emphasis on following rules, level of freedom employees are allowed in performing their duties, level of decision-making autonomy (Damanpour, 1991)
- Alignment with mission and goals of the organization: Crossan and Apaydin (2010)
- Organization commitment: includes support from the top, recognition of innovation efforts, resources for innovation (Borins, 2001) senior management support, funding, innovation champions, revised internal processes, metrics and incentives (Chesbrough and Crowther, 2006).

Project Level:
From observation and empirical analysis of all the projects hosted on the platform to date, distinct differences have been observed along certain dimensions (discussed later in preliminary findings). However it is not known what the effect of these differences have on the decision to implement and it is proposed that they be investigated further:
- Use of intermediaries: private contractors or external partners who participate in the planning or execution of the contest either through hire or mutual collaboration
- Type of Task: Varies based on the objective and complexity of the task at hand

Individual Level:
- Employee attitudes (Not-Invented-Here Syndrome): refers to internal resistance to outside knowledge (Chesbrough and Crowther (2006); Chesbrough (2011); Huston and Sakkab (2006)).
- Employee perception of benefits: innovations stand a chance of not being accepted if employees do not see the benefits. A similar concept (relative advantage) is used in the Diffusion of Innovations literature (Rogers,1995) signifying the perceived benefits of the innovation over the previous one that it is replacing.

Research and practice on open innovation and crowdsourcing in the public sector are emerging, and some of the influencing factors may not be identified in existing literature. The propositions above therefore form a starting point for the investigation, and it is expected that other factors will emerge from the empirical data collected.

Methodology
The investigation took the form of a multiple case study examining 28 open innovation projects across 17 agencies utilizing the challenges and prizes format. Stratified purposeful sampling was used to ensure proportional distribution of cases based on the type of task being crowdsourced. For each case, semi-structured interviews were conducted with one or two persons responsible for the administration or execution of the project. Interviews elicited responses about the decision to use the challenge format and organizational processes surrounding use of this approach. A combination of deductive and inductive analysis has been employed to identify the main enablers and barriers to practicing open innovation activities. To date some preliminary analysis has been conducted on the interview data which will be followed by a combination of more in-depth qualitative and content analysis. Maintaining a chain of evidence in addition to member checking is being employed to ensure accuracy of findings.

Preliminary Findings
The first phase of empirical analysis involved categorization of the 143 contests hosted on the platform from inception in September 2010 to December 31 2011 based on attributes such as type of task, objective, and target audience. The list of contests was later updated to include those posted up
until August 2012. One of the more striking findings was the level of disparity in the types of tasks targeted. For example, a large number seemed to revolve around public engagement with messages related to awareness and education campaigns. Many of these challenges had minimum requirements on specialized skills and the task could be accomplished by members of the general public. On the other end of the spectrum there were some highly technical tasks requesting submission of proposals or prototypes of the operational solution to specific problems. The target audience was generally individuals and groups with very specialized training or skills. Also noted was that a large number of contests employed third-party intermediaries to run or host the contest, while others were run exclusively by the host agency.

The following represents a preliminary categorization of tasks hosted on the platform:

i. *raising awareness of public services and issues* – included mainly creativity contests such as video, poster, and slogan contests aimed primarily at helping agencies spread the word on a particular issue.

ii. *providing a technical or tangible solution to a problem* – sought proposals, designs, prototypes, models leading to the creation of an actual product (e.g., energy saving light bulb, combat vehicle).

iii. *providing tools or methods to facilitate/improve provision of government services* – mainly solicited development of software applications to help citizens access government services.

iv. *generating research on a particular topic* – submission of white papers, conference papers, education initiatives, which made a general contribution to knowledge on a topic without necessarily tackling a specific problem.

The second phase of analysis involves deductive coding of interview data based on the proposed framework, as well as inductive coding to allow emergence of new factors. Analysis is focused on identifying determinants which influence the decision of an agency to use an open innovation approach to solve a problem. The analysis is currently at a very early stage and is expected to continue over the next few months. Preliminary results have so far provided support for some of the factors suggested in the literature such as the importance of management support, organization commitment, nature of the problem or task, and perception of benefits. Other determinants emerging inductively from the data so far include: origin of initiative (top-down vs bottom-up), and presence of a project champion.

**Limitations and Contributions**

With the platform barely two years old, current users may be viewed as early adopters. At this point it would be impossible to predict whether current trends being investigated will hold true after more years of operation.

The expected contributions include: i) development of a framework to represent enabling conditions for open innovation practices in public sector agencies ii) taxonomy of developed to represent the various dimensions of implementation. iii) recommendations on what constitutes an enabling environment including conditions under which it is most appropriate. In addition, implementers and administrators of the platform have indicated a keen interest in the findings and recommendations.

**References**


How Information Science Professionals Add Value in a Scientific Research Center

Christopher Eaker  
University of Tennessee  
ceaker@utk.edu

Andrea Thomer  
University of Illinois Urbana-Champaign  
athomer2@illinois.edu

Erica Johns  
University of Tennessee  
ejohns3@utk.edu

Kayla Siddell  
University of Tennessee  
ksiddell@utk.edu

Abstract

In response to the increasing need for a data curation workforce, the Data Curation Education in Research Centers (DCERC) program is educating library and information science students in scientific data curation. During the summer of 2012, the authors worked alongside scientists and data managers at the National Center for Atmospheric Research in Boulder, Colorado, to learn data curation within the context of a research center. Each student was matched with a “Science Mentor” and a “Data Mentor” based on prior work experience and the results of a placement questionnaire completed before the internship. Though NCAR has robust data services, we found that there was nevertheless still a role for data curators who can foster close collaborations between scientists and repository managers that may have traditionally not existed. This collaboration supports an otherwise impossible mutual education that benefits all involved. This poster demonstrates tangible outcomes of these internships.

Keywords: data curation, data management, metadata, data citation, data life cycle

Introduction

The volume of scientific data is growing exponentially across all scientific disciplines (Lynch, 2008; Hey & Trefethen, 2003). In response to the increasing need for a data curation workforce, the Data Curation Education in Research Centers (DCERC) program is educating library and information science students in scientific data curation through a partnership between the iSchools at the University of Illinois (Illinois) and the University of Tennessee (UT) and the National Center for Atmospheric Research (NCAR) (Palmer, Allard & Marlino, 2011). The project, funded by the Institute for Museum and Library Services, is supporting three master’s students from UT and three doctoral students from Illinois with diverse backgrounds in science and research. The students brought LIS perspectives and expertise in areas including information organization, user communities, and long-term preservation to the internship teams at NCAR (Palmer, Renear & Cragin, 2008), as well as skills obtained from prior careers. Their projects demonstrated the valuable roles that LIS data curation can bring to a large-scale data center setting, especially as intermediaries who foster collaboration between all project stakeholders.

Background

During the summer of 2012, four students (three from UT and one from Illinois) worked alongside scientists and data managers at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, to learn data curation within the context of a research center. NCAR has long-standing, sophisticated data services (i.e., extensive IT support staff, data repositories and repository managers), providing an ideal environment for students to learn current best practices in data management in a state-of-the-art research environment. The students brought LIS perspectives and expertise in areas including information organization, user communities, and long-term preservation to the internship teams at NCAR (Palmer, Renear & Cragin, 2008), as well as skills obtained from prior careers. Their projects demonstrated the valuable roles that LIS data curation can bring to a large-scale data center setting, especially as intermediaries who foster collaboration between all project stakeholders.


Copyright is held by the authors.
The Projects

Each student was matched with a “Science Mentor” and a “Data Mentor” based on prior work experience and the results of a placement questionnaire completed before the internship.

**Chris Eaker** (Data Mentor, Scot Loehrer; Science Mentor, Kate Young). The project addressed the following areas of the data life cycle: Collect, Describe, Preserve, and Analyze. With guidance from his mentors, Eaker conducted a data management audit within the In-situ Sensing Facility (ISF) in NCAR’s Earth Observing Laboratory (EOL). The audit identified areas of strength, areas needing improvement, redundant activities, and useful tools within groups’ data management practices. The goal of the audit was to increase efficiency and improve data management practices within the research groups, which will, in turn, support the long-term preservation and access of data sets. He produced a report with recommendations for improvements to ISF’s data management workflows in areas including data management planning, metadata, archiving and preservation, and project tracking.

**Erica Johns** (Data Mentor, Robert Dattore; Science Mentor, Dr. Samuel Levis). The project addressed the following phases of the data life cycle: Plan, Collect, Assure, Describe, Preserve, Discover, and Integrate. With guidance from her mentors, Johns accumulated, reformatted, and ingested scientific data from the Ameriflux network for an NCAR’s climate scientist, all the while seeking to create a valuable data product that would be accessible for future users through NCAR’s Computational and Information Systems Laboratory (CISL) Research Data Archive (RDA). ([http://bit.ly/PH8STV](http://bit.ly/PH8STV)). Johns then compared the data curation life cycle as she experienced it to the published DataONE model (Figure 1).

**Kayla Siddell** (Data Mentor, Steve Worley; Science Mentor, Dr. Patricia Romero-Lankao). The project addressed the following phases of the data life cycle: Plan, Collect, Assure, Describe and Preserve. Siddell acted as a translator and facilitator between a social scientist and repository manager by curating the first cross-disciplinary data set into the RDA. The final deliverable was a curated and archived data set, accessible on the RDA website ([http://bit.ly/S6qvZf](http://bit.ly/S6qvZf)) and a report on the lessons learned over the seven week internship containing data management tips for the research team.

**Andrea Thomer** (Data Mentor, Gary Strand; Science Mentor, Dr. David Schneider). The project addressed the following phases of the data life cycle: Describe, Discover, and Integrate. Thomer and her
mentors broadly explored ways of fostering a culture of metadata creation and data citation among researchers, and specifically explored ways of incentivizing contributions to the Climate Data Guide (CDG) – the "go-to source for scientifically sound information and advice on the strengths, limitations, and applications of climate data" (climatedataguide.ucar.edu). They added easily copy-and-pasteable "suggested citation" text to the top of each page in the CDG, in which contributors to the guide are listed as authors. They hope that by making the CDG a more formal, citable publication, scientists will be encouraged to make contributions. Additionally, they looked at metadata use in climate model output data to assess how often researchers made use of existing metadata standards. A report was written containing their findings.

Discussion

Though all four of these projects focused on different aspects of data curation and different portions in the data curation life cycle, they were similar in that the interns consistently acted as intermediaries between different groups in one of the NCAR research centers. Siddell, for instance, acted literally as a translator between the scientist and the repository manager, working closely with Romero-Lankao to first understand her data set, which was often written in Spanish and, in some cases, Spanish abbreviations, and then with Worley to understand how to best prepare the data set for ingest into the RDA. Similarly, Eaker’s project demonstrated the value of collaboration among researchers within the different research groups in the same laboratory. His project aimed to bring together researchers from separate groups to share tools, data management practices, and metadata to improve efficiency and aid in long-term preservation of data sets. Finally, Johns’ project required continuous collaboration with the scientists using Ameriflux data, the data provider, and the software engineer who made the final reformat and ingestion possible. She repeatedly appraised the usefulness and appropriateness of the data obtained for integration into one data product that would not only meet the current scientists’ needs, but would also be a valuable data product for future users.

In addition to serving as intermediaries, Information Science professionals trained as data curators are particularly knowledgeable about metadata and can assist researchers with describing their research data completely. Metadata allow scientists to make sense of data, i.e., how they were collected, why they were collected, who collected them, and how they were processed (Michener & Jones, 2012). Thomer found that scientists needed to be incentivized to add appropriate metadata to their data sets at multiple points in the data life cycle -- not just at the point of data creation, but also after first use. Eaker’s project confirmed what Tenopir, et al., discovered: researchers often do not use a metadata standard at all or use one that is specific to their laboratory (2011).

Conclusion

Though NCAR has robust data services, we found that there was nevertheless still a role for data curators who can foster close collaborations between scientists and repository managers that may have traditionally not existed. This close collaboration supports an otherwise impossible mutual education. The internships helped data managers on the team learn more about user needs, and researchers on the team gain a better and stronger understanding of the importance of good data management. Furthermore, in addition to the tangible outcomes of the students’ projects, the scientific researchers benefit by gaining an understanding of the importance of metadata at all steps during their project.

References


Using Design Thinking to Empower Ethnic Minority Immigrant Youth
in their Roles as Information and Technology Mediaries

Karen E. Fisher
University of Washington
Information School
fisher@uw.edu

Phil Fawcett
Microsoft Research
philfa@microsoft.com

Abstract

In the United States, almost one in nine people are foreign born (2010 U.S. Census). While diversity brings incredible richness, it also belies responsibility for helping immigrants participate fully in American life and culture versus isolating themselves within ethnic communities.

Our past work funded by the U. S. Institute of Museum and Library Services and the Bill & Melinda Gates Foundation shows two-thirds of people use library computers to search for information on behalf of someone else (US Impact Study, 2011). These Information Mediaries (Infomediaries) tend to be young, non-white, non-native English speakers, suggesting ethnic minority youth are key for understanding the information needs of their respective populations, passing on needed information to otherwise hidden users (especially non-English speaking, non-library users).

Our current research is over two phases supported by Microsoft and the Institute of Museum and Library Services we are employing design thinking (Brown, 2008, HBR, 2008), to study how ethnic minority youth serve as technology and information mediaries within their social networks. Our approach brings in perspectives from computer science, information science, design, information behavior, and uses research methods such as participatory design, action research, social networking analysis, rapid prototyping, and engineering. Our method integrates these perspectives and methods into a multi-day workshop format called “teen design days.”

Keywords: design thinking, teen design days, immigrants, infomediaries, cultural probes

In the United States, almost one in nine people are foreign born (2010 U.S. Census). While diversity brings incredible richness, it also belies responsibility for helping immigrants participate fully in American life and culture versus isolating themselves within ethnic communities. However, the needs for help with everyday life, particularly through effective information services and technologies, are acute due to an influx of immigrants from new countries, settlements outside traditional gateways, and increasingly complex language, technological, cultural, social, physical and financial barriers.

Our past work funded by the U. S. Institute of Museum and Library Services and the Bill & Melinda Gates Foundation shows two-thirds of people use library computers to search for information on behalf of someone else (US Impact Study, 2011). These Information Mediaries (Info-Mediaries) tend to be young, non-white, non-native English speakers, suggesting ethnic minority youth are key for understanding the information needs of their respective populations, passing on needed information to otherwise hidden users (especially non-English speaking, non-library users), and identifying the greater range of how libraries help.

In our current work funded over two phases supported by Microsoft and the Institute of Museum and Library Services we are employing design thinking to study how ethnic minority youth serve as infomediaries within their social networks. Design thinking originated from the design discipline and was coined by Dave Kelly and Tim Brown of IDEO in 2001. It was later popularized in a Harvard Business Review article by Tim Brown (HBR, 2008) when IDEO was called to solve a variety of problems outside the traditional realm of most design firms. Since then, design thinking has been used to solve problems in...
business and product design spaces. Design thinking involves 3 major steps: Inspiration, Ideation, and rapid prototyping (Brown, 2008).

Our approach brings in perspectives computer science, information science, design, information behavior and uses research methods such as participatory design, action research, social networking analysis, rapid prototyping, and engineering. Our method integrated these perspectives and methods into a multi-day workshop format called “design days.” These workshops are designed to gain an understanding of how and why immigrant and refugee youth act as infomediaries to members of their social networks, especially their ethnic communities (and elders), and how they can be supported through technology and services.

The core theoretical framework used in the design day workshops is based on a modified version of Meyers, Fisher and Marcoux’s (2007) Tween Day Methodology to understand how:

1. Ethnic minority youth surface the needs of others, communicate/provide information and technical help, and how they create, remix and manage information;
2. How elders and others engage with youth and benefit from provided assistance; and,
3. How these behaviors can be supported by design thinking (Brown, 2008, 2009).

The Teen Design Days are being held at near-by community centers. These centers are gathering places or “information grounds” (Fisher, Landry & Naumer, 2007) for many immigrant communities across the United States. At the Teen Design Days, participants reflect on their info medieary behavior using social network mapping tools, rapid prototyping, and cultural probes (Gaver, et al., 1999; Graham, et al., 2007) which involve storytelling, images, and dramatic play, as well as design, prototyping, and programming principles. Smart phones, laptops, Kinect controllers, and other devices are also being used along with technology experts from the University of Washington and Microsoft Research which together devise ways of facilitating teens’ current and future infomediary behaviors. At the our Teen Design Days, community members are meeting the teens’ needs for physical activity, competence and achievement, self-definition, creative expression, positive social interaction, structure and clear limits, meaningful participation in cross-culturally and gender appropriate ways through African dance class, beading, cooking, drumming, athletics, etc. The first phase of our Teen Design Days ends with a community showcase where the participants share their experience with their parents, project funders, instructors, and city staff, who select prototype designs for further development and implementation.

Phase Two broadens the use of design thinking and draws on the findings from the innovative Teen Design Day method honed in Phase One. We are carrying-out a mixed-method investigation with immigrant and refugee youth in the Seattle area that focuses on public library-community agency innovation. According to Burke (2008), of the 9,214 public library systems in the United States, about 75% serve patrons from over 30 countries, some up to 200 different nationalities. Libraries are thus poised to provide solutions and local leadership that can benefit entire immigrant communities and affect policy making and provide social and economic development. Our Phase Two research thus focuses on empowering immigrant youth through information literacy, computer and Internet services and other innovations at public libraries.

Using mixed methods, the research goals of our second phase are being accomplished over three stages:

1. Broad-Scope In-person Survey with 500 Seattle youth (ages 14-18; foreign-born, first generation immigrant, or speaking non-English language at home) to generate broad understanding of the nature of immigrant teen information mediary behavior, and,
2. Teen Design Days with 45 youth (groups of 15, minimum 3 locations, 3 occasions each) based on modifications to the techniques honed in Phase One and include providing feedback on paper prototyping of info mediary services. In some cases the prototypes generated from Phase One will be field tested; and,
3. Workshops with public library and community organization staff on designing services for info mediary ethnic minority youth. Project partners include the Seattle Public Library, the King County Library System, Horn of Africa Services, Vietnamese Friendship Association, and the YMCA of Greater Seattle.
Our novel research approach to understanding the relationships between information, technology, design, and empowerment in immigrant youth populations, who act as information mediaries, is poised to make significant future contributions to the fields of information science, computer science, human computer interaction, and design.

References


Old Maps and Open Data Networks

Bernhard Haslhofer, Keith Newman, Amanda Stefanik
Cornell University, Information Science, Ithaca, NY
{bh392, kin7, ams568}@cornell.edu

Werner Robitza
University of Vienna, Austria
werner.robitza@univie.ac.at

Carl Lagoze
School of Information, University of Michigan
clagoze@umich.edu

Abstract

Old maps are a record of the past, exposing features people might want to tell stories about. Maphub is a Web application that enables them to do so by creating annotations on digitized high-resolution historical maps. By semantically tagging regions on the map, users create associations between their annotations and resources in open Web-based data networks. These associations are leveraged to enable multilingual search and to generate overlays of historical maps on modern mapping applications. Contributed annotations are shared on the Web following the W3C Open Annotation specification. Preliminary studies show general user satisfaction with our approach.

Keywords: historical maps, annotations, linked data, GIS

Introduction

Historic maps reflect geographic information and also the attitudes, perspectives, and beliefs of different times. Their geographic accuracy tells us much about the state of geographic knowledge and technology at the time of their creation. Thousands of maps have already been converted to digital form and made available online and accessible to scholars and the broad public (Rumsey & Williams, 2002).

When viewing digital maps users often have stories to tell. They might know something about the context of a map, might be able to identify places or landmarks, or might have comments that could complement historic maps as records of the past. However, most current map hosting environments don’t allow users to contribute their stories.

We believe that allowing annotations on historic maps is a possible way of recording these stories. If we also connect named entities mentioned in these stories with other related resources in open globally connected data networks, such as DBpedia (Auer et al., 2007) or GeoNames, we can capture the context of these stories and, through named entity disambiguation, we can connect the stories of different individuals. This user-contributed information, in combination with curated metadata, is a valuable source for search and retrieval or any other data analysis task.

The Maphub prototype is the result of a demonstration experiment carried out as part of the Open Annotation Collaboration (Hunter, Cole, Sanderson, & Van de Sompel, 2010). It showcases how users can annotate historic maps and connect these annotations with resources in open data networks via a function, called semantic tagging. Tagging creates relationships between annotations and Web resources. These relationships are leveraged to supplement user-contributed annotations with additional information from the Web, enabling functions such as multilingual search and retrieval over historic maps. Collected annotation data are contributed back to global data networks by exposing them as dereferenceable Web resources following the W3C Open Annotation specification.

References

1. http://maphub.github.com
2. http://www.w3.org/community/openannotation/


Copyright is held by the authors.
The Maphub Prototype

The Maphub prototype is a Web portal, primarily based on Ruby on Rails and JavaScript. It allows users to tell their stories about areas of interest by letting them annotate regions on historic maps. A first demo, which has been bootstrapped with approximately 6,000 high-resolution digital maps from the Library of Congress Map Division’s catalogue, is available at http://maphub.herokuapp.com/. Users can browse maps or search for topics of their interest. Once a map is selected, the user can explore it by panning and zooming using either the mouse or keyboard. The current release (v0.3) implements two main annotation tools, one for commenting on map regions and one for georeferencing known locations on a map.

Using the commenting tool (“Annotate”), it is possible to draw shapes on the map, such as rectangles, lines and polygons (Fig. 1). These shapes mark a region of interest, and once the user has finished drawing, a popup will ask them to enter their comment and tell their story about that region in a free text field.

![Figure 1. Polygonal annotation with tooltip](image1)

While the user is writing text, the Maphub system analyzes the input and proposes possibly relevant semantic tags. These tags are links to Web resources such as Wikipedia articles and are suggested by querying open data sources such as Wikiminer or GeoNames (Fig. 2).

![Figure 2. Full annotation with semantic tags shown below the map](image2)

To disambiguate the meaning of these tags, each tag carries a short description of its subject, available as a tooltip. Furthermore, we allow the user to accept or reject individual tags before annotations are being saved. In the background, Maphub then dereferences the URIs of accepted semantic tags and adds subsets of the retrieved data representations to the index. This way users can, for example, search for maps in any language supported by Wikipedia.

The georeferencing tool ("Control Point") allows users to associate known locations on a historic map with location resources provided by the GeoNames web service. As soon as we have at least three control points, we can compute a model for translating between the x/y pixel dimension of the digitized historic map image and real-world projections used in modern mapping systems. This allows us to visualize historic map overlays in Google Maps and project maps onto a three-dimensional globe, provided by Google Earth (Fig. 3).
Sharing collected annotation data in an interoperable way was another major goal of this demonstration experiment. Maphub is an early adopter of the Open Annotation model, which is currently specified in the W3C Open Annotation working group. It demonstrates how to apply that model in the context of digitized historic maps. As described in the Maphub API documentation\(^3\), each annotation becomes a first class Web resource that is dereferencable by its URI and therefore easily accessible by any Web client. In that way, while users are annotating maps, Maphub not only consumes data from global data networks—it also contributes data back.

**Preliminary Results and Planned Experiments**

The focus of the first project phase was on the design and implementation of the previously mentioned annotation functions on a real-world historic map collection. A major goal was to apply the Open Annotation specification and to provide feedback for its further development. The main observations made were that the model can easily be implemented and that is expressive enough to describe map annotations in an open, Web-based, and interoperable way. However, the conceptual simplicity of the supported annotations and the relative technical complexity and verbosity of the resulting annotation serializations raised some concerns on the model's usability.

We conducted first usability tests, which showed that users generally accept and understand the notion of annotations in the context of historic maps. We were able to identify issues to be resolved regarding the usability of the user-interface.\(^4\)

We also prepared the setup for an experiment to study the possible effects of the semantic tagging function on user behavior and therefore also on the outcome of the annotation process. The experiment follows an in-lab within-subject design with varying tagging conditions and will be the main focus of our future work.

**Summary and Next Steps**

Maphub is a Web-application that allows users to tell stories about regions of interest on historic maps. While they create annotations, the system supports them in associating their annotations with other related Web resources by proposing semantic tags. This contextualizes the stories being told and establishes a source of additional data, which can be used for tasks like multilingual retrieval or enhancements such as map overlays. In that way, annotations are contributed back as data nodes in a global data network.

Currently we are conducting an in-lab, within-subject design experiment, with the goal of learning about the effects of the semantic tagging function compared to other established tagging mechanisms such as free tagging. Afterwards we would like to release a first public version and ask possibly interested users to participate in a first beta-release stage. In parallel to these experiments, we are seeking feedback from users and institutions and will use this as input for further developments.

---

\(^3\) [http://maphub.github.com/api](http://maphub.github.com/api)

\(^4\) [https://github.com/maphub/maphub-portal/issues](https://github.com/maphub/maphub-portal/issues)
References


Literature Review: How Content Analysis Is Used to Study Human Rights Violations in Political Science?

Tatiana Vashchilko  
Faculty of Information and Media Studies  
University of Western Ontario  
tvashchli@uwo.ca

Lu Xiao  
Faculty of Information and Media Studies  
University of Western Ontario  
lxiao24@uwo.ca

Abstract

Our study focuses on the review of the existing datasets on human rights violations to understand the types of data sources used to construct the datasets and the methodologies to measure the degree of human rights violations in different countries. This poster paper reports, preliminary findings from Political Science research in 1971-2012 on the scope of employed human right violations’ data and methodologies.

Keywords: human rights, content analysis, political science

Introduction

Quantitative analysis of human rights violations is important for systematically testing existing theories explaining human rights violations along with their consequences, and design of effective policy instruments to prevent future human rights violations. Amnesty International and The US Department of State’s Human Rights Reports were the first ones in the early 1980s to provide any systematic textual description of human rights violations across different countries on an annual basis. Later on, these reports have become the basis for two-three large-scale datasets with the quantitative measures of the degree of human rights violations across multiple dimensions (e.g., civil rights, political rights). As the importance of reporting human rights violations by governmental and non-governmental organizations rose, the primary sources with the rich textual data offering additional details about the victims, perpetrators, and events have emerged. With the recent availability of primary information on human rights violations, and advances in natural language processing (NLP) to quickly identify and extract systematic data from textual information, this paper analyzes how scholarly community have been utilizing evidences of human rights violations. The paper contributes to research on human rights violations by systematically accounting for the existing use of primary and secondary data on human right violations in Political Science, one of the main disciplines studying this phenomenon. In this poster paper, we report preliminary findings on the identification of the majority of human rights data sources employed in Political Science based on the analysis of peer-reviewed articles in 1971-2012. The paper deliberates first on the corpus, then proceeds with the description of the coding scheme, and finally discusses the results and implications.

Corpus and Coding Schema

There are two major databases in political science research: Worldwide Political Science Abstracts (WPSA) operated by ProQuest and International Political Science Abstracts (IPSA) operated by Ovid Wolters Kluwer. We chose WPSA in our study because it covers larger period of publication time (1923 – present), and offers greater journals’ coverage. For example, searching “human rights” phrase in IPSA produced only 5,412 results, peer-reviewed publications, whereas WPSA produced 21,949. Because our goal is to understand the use of human rights violation data in the research studies and how they were analyzed, we used filters such as “data”, “analysis,” “case stud,” “empiric,” or “statistic” to further specifying the literature body we are interested in. With only peer-reviewed English articles included, our search produced 2,796 items published in 1971-2012 (as of July 30th, 2012). We then used two criteria to
select the papers for further analysis. The first and most important one is that the reported study should have used evidence on human right violations in the analysis (e.g., testimonies of victims, photographs, etc). The second criterion intended for including each of the quantitative datasets on human rights violations only once in our analysis papers and, therefore, had two requirements: 1) the quantitative datasets of human rights violations had to be constructed by the authors of a peer-reviewed article under consideration, and 2) the measures of human rights violations in the quantitative datasets had to be based on textual or non-textual human rights violation documents. Including only these types of peer-reviewed publications advanced the understanding of content analysis used to extract quantitative characteristics of human rights violations from textual and non-textual sources. After these criteria were applied, an expert opinion of a political scientist, one of the authors of this paper with a doctorate degree in Political Science, was used to determine whether a paper should be included or not based on these criteria by reading through the titles, abstracts, and sometimes full texts of the papers, if needed, of all selected 2796 peer-reviewed articles. After this first step, 152 potential articles from Political Science journals were selected for the subsequent more nuanced analysis. Finally, after careful consideration of 152 articles, 31 peer-reviewed articles published in Political Science were included in the corpus to conduct the analysis of the utilized sources on human rights violations. To make sure that our selection is comprehensive, we compared our generated list with the reference lists of major literature review papers such as Green (2001) and Milner et al., (1999). After the extensive literature search, careful identification and subsequent verification processes, the belief is that the corpus includes the vast majority of peer-reviewed articles in Political Science employing at least one information source on human rights violations.

We coded the selected 31 articles focusing on the information sources and data analysis. The example codes include characteristics of the sample (e.g., sample size, year coverage), data source, data accessibility, data format, and the kinds of data analysis methods. We also coded the software programs used in the analysis, if any.

**Preliminary Findings**

We found that the papers in the sample were predominantly published in the last five years (Figure 1). Apart from the growing scholarly interest in the issues of human rights abuses, there seems to be two main reasons that drive this trend: 1) the availability of information on human rights violations, and 2) the advances in software development for content analysis. For example, the need for collection and preservation of information on human rights violations has been brought to light with the potential of better identification of victims and predators of past crimes if employing systematic qualitative and quantitative analysis (Rosenberg, 2012).

![Figure 1. Distribution of Papers in Political Science by the Year of Publication](image-url)
In terms of qualitative analysis, we considered three modes: manual, semi-automated, and automated. Both the hand coding and the use of software programs like NVivo are considered manual coding. Two studies used algorithms that take coding schema and corpus as input and generated coded results as the output. For example, Chapman (2007) uses semi-automated analysis to evaluate the experience of the South African Truth and Reconciliation Commission (TRC) and assess whether the TRC was successful in developing a model to promote intergroup forgiveness and reconciliation. In particular, to understand the full range of victimization experience, the transcripts of 429 victims were quantitatively coded using ATLAS/ti computer-based program. No study used completely automated approach for content analysis, that is, the use of an algorithm that takes corpus as the input and generates coded results as output with natural language processing and text mining mechanisms.

Most of the papers were using publicly available sources of information (Figure 2). Also, 17 studies relied on secondary data, 12 on primary data, and 2 on both. We found that if sample included many countries and time periods the studies tend to rely on secondary and public data. The “two most widely distributed and read sources on information on countries’ human rights practices” are the US Department of State Country Reports and the Amnesty International annual reports (Poe, Carey, & Vazquez, 2001, p. 653). Most of the papers on human rights violations used the event-level or country-year level of analysis (Davenport & Ball, 2002).

The majority of the papers, which used primary data and private information, focused on one country or two countries to investigate predominantly the specifics of human rights violations committed in those countries. However, in general, very little primary documentation on human rights violations is available in research community. Montgomery (2001) pointed out four reasons that contributed to this: 1) NGO’s concern about potential access to the victims’ testimonies by the perpetrators; 2) inability of libraries and archives to recognize the importance of preserving human rights violations’ documentation; 3) financial constraints of existing archives of human rights violations documents; and 4) reluctance of private foundations to fund the human rights violations archives.

Discussion and Future Work

Political science developed two main approaches to measure human rights violations. One way to measure human rights violations is by means of coding cases of repressive events from media, so-called events-based approach (Milner, Poe, & Leblang, 1999). Another one is standards-based approach, which measures the degree of human rights violations in different countries by coding governments’ reports on human rights and categorizing countries according to a “predetermined criteria” (Milner et al. 1999, p.
For example, a very well-known measure of human rights violations, Political Terror Scale (PTS), was developed in the 1980s (Stohl & Carleton, 1985), and has been used by many researchers in Political Science to code various textual reports on human rights violations across different countries (Gibney, Cornett & Wood 2012).

The standards-based approach is the preferred one in Political Science due to lesser extent of potential biases and greater comparability of degree of human rights violations across countries. However, it sacrifices the details of the human rights violation events. Sylva and Wyly (2001) demonstrated the emerging benefits of the events-based approach investigating the nature of coverage of Sierra Leone civil war in the Western media, specifically in New York Times, Sylva and Wyly (2001) identified that “the single most frequent reference in the articles analyzed was to the types of atrocities committed or alleged. Almost half provided information on human rights violations, and more than one-third included estimates of the number of war casualties” (p. 45). Thus, Sylva and Wyly (2001) demonstrated that a major western newspaper reported on human rights violations in a detailed manner. With globalization and the spread of the Internet, the events-based approach based on the media coverage of various military conflicts might generate improved quantitative human rights violations indicators, which can, if not supersede, at least be comparable to the standard-based approach and depending on the aggregation of the media outlets, less biased than the standard-based approach. This implies the need of developing algorithms for analyzing the amount of data introduced through such approach. It is also expected that the availability of such tools for analyzing big data may affect how human rights violation studies are conducted in the future. The development of such algorithms is the ultimate goal of this funded project (funding agency: Social Sciences and Humanities Research Council in Canada).

References


Research Specialties as Emergent Phenomena: Connecting Emergence Theory and Scientometrics

T. Clay Templeton  
University of Texas at Austin  
clayt@ischool.utexas.edu

Kenneth R. Fleischmann  
University of Texas at Austin  
kfleisch@utexas.edu

Abstract

This short paper represents an initial effort to connect the emergence theory literature with the bibliometric, informetric, and scientometric literatures. It begins with a prominent definition of emergence, and then connects each of the components of this definition with the relevant insights about the development of new scientific and technical concepts or research specialties. Finally, it concludes with a discussion of the relationship between these two distinct areas of scholarly inquiry and the need for further exploration of this intersection.

Keywords: emergence theory, emergent phenomena, bibliometrics, informetrics, scientometrics, citation analysis, cyberinfrastructure

Introduction

Emergence refers to a situation in which a qualitatively novel entity, or an emergent, is generated through the interactions of lower-level entities (Sawyer, 2005). Emergence theory thus seeks to describe a wide range of physical, biological, and social phenomena. Since science can be viewed as a social phenomenon (Fleck, 1979), the development of new scientific and technical concepts and research specialties can be seen as a process of social emergence (e.g. Guo et al. 2011, Leitz 2009, Chen et al. 2009). As emergence theory has organized our thinking about the social lives of scientific specialties and concepts, the literature these specialties and concepts left behind beckoned as a realm of data in which to explore its workings. In this paper, we recount this interaction by mapping ideas from a prominent definition of emergence (Goldstein 1999) to ideas in the bibliometric, informetric, and scientometric spaces.

What Is an Emergent Phenomenon?

Goldstein’s (1999) work can be seen as part of the third wave of social systems theory, which situates social emergence within a broader framework of interest in systems whose evolution is sensitive to environmental conditions (Sawyer, 2005). Goldstein’s focus is on organizational dynamics, and thus, his characterization of emergence is highly compatible with an exploration of science as a social phenomenon. According to Goldstein’s definition, emergent phenomena are characterized by five characteristics:

Acknowledgements: This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20153. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.


Copyright is held by the authors.
• “Ostensive: Emergents are recognized by showing themselves, i.e., they are ostensively recognized.” (p. 50)
• “Global or macro level: The locus of emergent phenomena occurs at a global or macro level. Observation of emergents is of their behavior on this macro level.” (p. 50)
• “Coherence or correlation: Emergents appear as integrated wholes that tend to maintain some sense of identity over time.” (p. 50)
• “Dynamical: Emergent phenomena are not pre-given wholes but arise as a complex system evolves over time.” (p. 50)
• “Radical novelty: Emergents have features that are not previously observed in the complex system under observation” (p. 50)

Ostensive

Emergents are recognized when they show themselves, which raises the question of to whom and under what conditions they become perceptible. This question is especially pressing for social emergence, since the ontological status of collective entities in the social sciences is controversial (Sawyer 2005). In the case of scientific concepts and specialties, how do we most definitively figure out which ones exist, what their proper names are, and where they begin and end? Arguably, scientific paradigms, fields, subfields, and specialties show themselves most nakedly to scientists working in or near those specialties. Practicing scientists are intimately acquainted with the details of professional scientific practice, and their professional practice depends on an ability to navigate the structure of science. They navigate, however, with the aid of bibliographic tools. Since the advent of citation databases (Garfield 1955), information professionals have augmented the perception of scientists and others with maps of science built from the ground up (Garfield et al. 1964, ISI 1981, Börner et al. 2003). The aim is to allow emergent scientific structures to show themselves earlier, in more detail, to more people.

Global/Macro Level

The locus of emergent phenomena occurs at a global or macro level, even when objective observation may be more easily achieved at a local or micro level. Scientific specialties and concepts emerge from a mangle of journals, authors, papers, keywords, and citations, among other things outside of bibliographic control (Pickering 1995). These basic units of bibliometric analysis are amenable to quantification (Borgman and Furner 2002), which makes them an appealing starting point for generating analysis of higher-level structures. Procedures for generating macro level structures from bibliometric data can be found in review literatures on mapping scientific specialties (Morris and van der Veer Martens 2008) and visualizing knowledge domains (Börner et al. 2003). The boundaries of scientific specialties (as constructed from bibliometric networks) remain open to interpretation, however. At present, information specialists frequently turn to Subject Matter Specialists (SMEs) to validate the results of mapping procedures (Morris and van der Veer Martens 2008).

Coherence or Correlation

Emergents appear as integrated wholes that tend to maintain some sense of identity over time; The entities involved in a research specialty tend to be more densely associated with each other than to entities outside the specialty. This basic assumption informs most approaches to extrapolating research specialties from bibliometric facts. For example, approaches based on co-citation analysis (Garfield 1979) first calculate similarity of papers based on how frequently each pair of papers appears together in reference lists, then cluster papers based on similarity. Approaches based on co-word analysis (Callon, Law, and Rip 1986) calculate similarity of words based on how frequently two words appear together in some part of a paper’s content or metadata, then cluster words based on similarity. Morris and van der Veer Martens review a number of approaches for constructing maps based on how bibliometric entities cohere in the written record of science (2008).
Dynamical

Emergent phenomena are not pre-given wholes but arise as a complex system evolves over time. The density of associations between entities in an emerging specialty tends to increase over time (Leskovec et al. 2005), and some authors have taken the emergence of a giant connected component in the coauthorship network as a signature feature of emergence (Bettencourt et al. 2009, Lietz 2009). In approaches to mapping scientific specialties based on clustering, a major challenge is tracking clusters from one time period to another. Asur et al. (2007) and Spilopoulou et al. (2006) provide equations for classifying changes in cluster structure over time that are agnostic with respect to clustering method. Research in literature dynamics (Tabah 1999) attempts to describe the processes by which specialties emerge using a variety of mathematical modeling techniques.

Radical Novelty

Emergents have features that are not previously observed in the complex system under observation. It may not be immediately apparent that a new specialty has emerged, and it is challenging to predict what the characteristics of a new specialty might be. Bibliometrics researchers have raised the question of how early a new specialty can be detected (Meadows and O’Conner 1971, Small 2006). Based on information foraging theory (Piroli 2007), Chen et al. (2009) suggest that an emerging research area might be signaled by citation bursts (Kleinberg 2002) to a paper that bridges existing areas of knowledge (Burt 2004). Ohniwa et al. (2009) describe a method of identifying emerging topics based on keyword increments from one time period to the next. Guo et al. (2011) incorporate word bursts, new author counts, and interdisciplinarity of citations (Porter and Rafols 2009) into their model of emergence.

Connecting Emergence Theory and Scientometrics

As we have tried to illustrate, the development of a new scientific or technical concept or a new research specialty as explored in the bibliometric, informetric, and scientometric research literatures can be viewed as a compelling case study of emergence. In turn, emergence theory helps to put these literatures into a broader context that potentially transcends science as an area of focus, by exploring the connections between science and other human, biological, and natural domains. It will be useful to further explore the connections between these two areas of research, in the interest of bridging this divide and hopefully furthering the research programs of both of these areas.

References


Digital Curation Tools: Metadata Enhancement with Selenium IDE

Daniel Gelaw Alemneh
University of North Texas Libraries
Daniel.Alemneh@unt.edu

Andrew James Weidner
University of North Texas Libraries
Andrew.Weidner@unt.edu

Abstract

Maintaining usable and sustainable digital collections requires a complex set of actions that address the myriad challenges at various stages of the data lifecycle. Digital curation activities enhance access and retrieval, maintain quality, add value, and facilitate use and re-use over time. Digital resource lifecycle management is becoming an increasingly important topic as digital curators actively explore tools and applications that directly perform curation and management tasks. Accordingly, the University of North Texas (UNT) Libraries develop and/or adopt various tools, workflows, and quality control mechanisms that enable quick and effective analysis and quality assurance. This brief paper demonstrates automated metadata enhancement with Selenium IDE, an open source, Web-based tool which UNT has adopted for use during the post-ingestion stage of the data lifecycle.

Keywords: digital curation, lifecycle management, metadata, open source, curation tools

Introduction

Digital lifecycle management starts when an item is created (born-digital) or selected for digitization (analog) and continues through image post-processing, metadata capture, derivative creation, and preservation for long-term access. Quality metadata is crucial to implementing reliable, usable, and sustainable digital libraries (Sumner & Custard, 2005). Recognizing the role of standardized metadata in digital resource lifecycle management, the University of North Texas (UNT) Libraries actively promote metadata-based digital resource management.

The UNT Digital Libraries Division utilizes various tools to ensure metadata consistency and precision across all digital resources and facilitate digital curation activities. This paper describes a workflow that uses Selenium IDE to edit large sets of published metadata records quickly and accurately with minimal human intervention.

What is Selenium IDE?

Selenium IDE is a free and open source add-on for the Firefox Web browser. It is primarily used by the Web development community to perform automated testing of Web applications. Selenium IDE provides an integrated development environment in which to create, debug and run custom scripts that automate actions in a Web browser. Users write or record scripts in the Selenium IDE window (Figure 1) and use standard play controls to run single scripts, called test cases, or groups of scripts, called test suites. The Selenese syntax, encoded as an HTML table, sends commands to the browser that act on specified page elements in sequence. See Table 1 for a list of common Selenium IDE commands.

Copyright is held by the authors.
Selenium IDE Workflow for Post-Ingest Metadata Normalization

Selenium IDE is an important part of the digital curation toolkit for the UNT Digital Libraries Division. It has proven to be essential in the ongoing processes of improving and maintaining metadata quality for large collections of digital objects. Naturally, as careful repository stewards, we attempt to publish accurate and complete metadata when we upload items to our repository. We have a number of tools at our disposal that facilitate normalized metadata creation and eliminate mistakes before items are uploaded.

Sometimes, for a variety of reasons that are usually out of our control, we end up with incorrect or sub-standard metadata for published digital objects. After objects have been published, our content management system provides a single method for editing metadata: the object record. A human operator must open an object record in a Web browser and manually change the information via the editing interface. When large sets of records contain metadata that must be normalized in order to improve retrieval or meet our repository's data standards, the single object method is undesirable.

The single object paradigm means that editing large sets of records requires shifting staff time away from more important production activities. However, if the metadata that must be changed is standard across the entire set, we use Selenium IDE to automate the editing process. In the best case scenario, a Selenium IDE operator creates a test suite that automates the editing process for multiple object records. If a test suite is not feasible, an operator implements a test case that streamlines the editing process for individual object records.

A typical metadata editing workflow begins with identifying a set of objects that require normalization. If the set is large and the required changes are standard, an operator creates a Selenium IDE script that performs the changes, publishes the new metadata, and closes the browser tab. After testing and debugging to ensure that the script performs correctly, the operator creates a test suite. Using the content management system's search interface, the operator opens multiple object records as Web browser tabs. Finally, the operator runs the test suite. Each script in the test suite works on a tab in the browser window until there are either no more scripts in the suite or no more tabs in the browser. The operator repeats the process until the entire record set is normalized.
Table 1
Common Selenium Commands

<table>
<thead>
<tr>
<th>Command</th>
<th>Function</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>assertValue</td>
<td>Tests an element's value.</td>
<td>Verify that a page element has a specific value. Accepts regular expressions.</td>
</tr>
<tr>
<td>click</td>
<td>Performs a mouse click.</td>
<td>Click links, buttons, and check boxes.</td>
</tr>
<tr>
<td>clickAndWait</td>
<td>Performs a mouse click and waits for the page to load.</td>
<td>Execute the next command in the script after the new page loads.</td>
</tr>
<tr>
<td>close</td>
<td>Closes the current tab or browser window.</td>
<td>Place at the end of a script to enable automated batch editing with a test suite.</td>
</tr>
<tr>
<td>keyPress</td>
<td>Sends a key press.</td>
<td>Delete or add characters in a text input element. Use with setCursorPosition.</td>
</tr>
<tr>
<td>select</td>
<td>Selects a value.</td>
<td>Select a choice from a drop-down menu.</td>
</tr>
<tr>
<td>setCursorPosition</td>
<td>Places cursor at a specific location in a text input field.</td>
<td>Use the variable with the type command to paste the assigned value in a text input field.</td>
</tr>
<tr>
<td>store</td>
<td>Assigns a value to a variable.</td>
<td>Use the variable with the type command to paste the assigned value in a text input field.</td>
</tr>
<tr>
<td>storeValue</td>
<td>Stores existing text in a variable.</td>
<td>Use the variable with the type command to paste the stored value in a text input field.</td>
</tr>
<tr>
<td>type</td>
<td>Enters specified text.</td>
<td>Populate a text input field. Overwrite existing text.</td>
</tr>
</tbody>
</table>

Use Case: The Portal to Texas History

The UNT Libraries (UNTL) metadata guidelines specify that the main title of an object is the title printed on the title or front page. We use a template tool to create standard metadata for large groups of similar objects, such as newspaper issues, during ingest. In some cases it is difficult to notice minor changes in a newspaper title during the quality control process. After the objects have been uploaded to The Portal to Texas History, however, we can easily view changes in a newspaper title by browsing through the thumbnail images created during the ingest process.

If we identify any title changes that are not reflected in the template-produced metadata, we use Selenium IDE to quickly bring our metadata in line with the UNTL standard. What follows is a step-by-step breakdown, with HTML code and a screenshot (Figure 2), of how a newspaper title script works:

1. Test that the original title is what we expect it to be with the “assertValue” command and a regular expression. Adding tests to our scripts ensures that we avoid making inadvertent changes to the metadata; the script will stop if the value does not match. If the title were the only information in the Main Title field, we would simply enter the title as the value. Because there is unique information present, such as volume and issue numbers, we use a regular expression to check the title from the beginning of the string.

   `<tr>
   <td>assertValue</td>
   <td>//div[@id='main']/div/div[2]/div/input</td>
   <td>regexp:^UNT Daily</td>
   </tr>`

2. Create a variable called “NewText” which contains the word “The” with the “store” command.

   `<tr>
   <td>store</td>
   <td>The</td>
   <td>NewText</td>
   </tr>`
3. Create a variable called “OriginalTitle” which contains the current text in the Main Title field with the “storeValue” command.

   `<tr>
       <td>storeValue</td>
       <td>//div[@id='main']/div/div[2]/div/input</td>
       <td>OriginalTitle</td>
   </tr>`

4. Paste the stored text from both variables, with a single space between them, into the Main Title field with the “type” command.

   `<tr>
       <td(type</td>
       <td>//div[@id='main']/div/div[2]/div/input</td>
       <td>${NewText} ${OriginalTitle}</td>
   </tr>`

5. Save the new metadata by clicking the Publish button with the “clickAndWait” command.

   `<tr>
       <td>clickAndWait</td>
       <td>name=publish</td>
   </tr>`

6. Close the tab with the “close” command to allow the next script in the automation suite to work on the next tab in the browser window.

   `<tr>
       <td>close</td>
   </tr>`

---

**Figure 2:** Newspaper Title Script in Selenium IDE’s Table View.

Combining multiple instances of the above script in a test suite automatically adds the word “The” to the beginning of the Main Title field for any number of digital object records loaded in browser tabs. In this manner we can quickly edit large sets of records and avoid the inevitable typographical errors introduced during manual data entry.
Conclusion

Large digital collections present challenges when producing descriptive metadata. Naming schemes and element definitions can vary widely, requiring substantial rework to meet local repository standards. Successful metadata enhancement strategies involve mechanisms for both pre- and post-ingest metadata normalization. Automated metadata normalization with Selenium IDE improves operator efficiency and accuracy during the time- and labor-intensive post-ingest data entry process (Figure 3).

From institutional repository platforms (e.g., DSpace) to commercial content hosting sites (e.g., Flickr), Selenium IDE can be used to edit metadata in any content management system that has a Web-based editing interface. Selenium IDE is a highly recommended addition to the metadata enhancement toolkit for any institution that serves content in a content management system with a Web-based administrative interface.

Figure 3: UNT Digital Library Workflow Modified for Post-Ingest Metadata Normalization. Adapted from “Metadata Quality Enhancement for Large Digital Collections: Web Browser Automation with Selenium IDE” by A. J. Weidner and D. G. Alemneh, 2012, UNT Digital Library.

References


Abstract

In the following article we examine how differences in location of @-mentions constitute both differences in meaning and social signaling in the context of the 2012 United States Presidential election. Through our analysis we establish that the Twitter handles of Barack Obama and Mitt Romney have distinctly different positions in tweets where they are mentioned. We expand on this finding by exploring how these positions change over time and how different syntactical features affect this position.

*Keywords:* Twitter, political discourse, social media, syntactical features

Introduction

Twitter is a global platform where participants dynamically construct context and express their positions on a wide range of topics. Over time, Twitter users have developed and adopted a set of prescribed syntactical features that to participate in conversations, exchange information and frame both durable and ephemeral context. The most widely used syntactical features include @-mentions, hashtags and retweets. The content and the resulting, dynamically constructed context depends on humans embedding technological symbols in the content of their message. Whether it is a retweet, an @-reply or @-mention, these symbols in the context of Twitter imply action possibilities.

To date, little research has been done on positions of syntactical features on Twitter (Zappavigna, 2011). We have been seeking to close this gap through analysis that spans multiple domains and multiple syntactic features. The research presented here examines @-mention position in a narrow context of political discourse in the United States; specifically, tweets referencing the two candidates running for office in the 2012 United States Presidential Election.

Literature

In Twitter, @-mentions are used to direct a tweet towards another user and insert one’s tweet into that person’s stream. Because @-mentions serve as a technologically specific hyperlink (a “learning affordance”) (Kaptelinin and Nardi, 2012) we conceptualize them here as a technological affordance. When considering @-mentions as affordances, they are technologically constitutive and only perceived when made. One must affix an @-mention to content for it to be viewed as a technological affordance.

In the case of @-mentions, we take the view of @-mention as both an affordance, in the technological sense, and its demonstrated role in coherent conversation and demonstration of one’s goals. Thelwall et al. (2011) propose applying the concept of *affordances* to understanding the purpose of function of tweets in terms of users goals during an event, a proposition that aligns itself with the uses and gratifications model of media, which argues that people actively exploit media for their goals. In Thelwall’s et al. (2011) conceptualization, events are more narrowly defined as an exogenous occurrence that results in an increased use of one or more words. Here we consider news pertaining to the incumbent President and opponent during an election season to constitute an ongoing event. We hypothesize @-mention use in the context of presidential candidates to be informed by a desire to demonstrate conversation (if in representation only) or insinuate talk. That is, its use merely implies conversation.
Dataset and Methods

As part of a larger collection effort examining the 2012 United States Election, we collected every time the handle of Barack Obama (@barackobama) and Mitt Romney (@mittromney), the two major candidates in the 2012 US Presidential Election, were mentioned on Twitter. This time period incorporated April 1, 2012 through August 13, 2012. Our data collection was conducted using the TwitterZombie architecture that queries the SEARCH API for tweets that meet specific selection criteria identified by the researcher (Black et al., 2012). We bound our data as this timeframe represents a distinct time in the United States where civic discourse shifted from the Republican Party primary season to the general election up through the time that immediately precedes the party conventions. Our analysis follows similar methods used in previous studies of political events on Twitter (Black et al., 2012; Mascaro and Goggins, 2012; Mascaro et al., 2012).

Research Questions

Our dataset represents a specific political context, but we believe it can be applied to other domains. We ask the following questions:

1. To what extent does the position of @-mentions of a national level presidential candidate vary?
2. What factors influence the positions of @-mentions for the two Presidential candidates?

Findings

Mention Position and Syntactical Feature Distribution

Table 1 represents the distribution of the syntactical features along with the median position of the handles of Barack Obama and Mitt Romney. We see that there were significantly more tweets collected with Obama’s handle. The median percent through each tweet that @BarackObama appears is 10.24%. In contrast, we see that @MittRomney occurred 21.77% of the way through a tweet, indicating that the handle occurred more towards the middle of the tweet. We also see that the distribution of other syntactical features was much more spread out with @BarackObama having a greater percentage of links and retweets and @MittRomney having more hashtags and @-replies. This difference is notable since we would expect that the earlier position of @BarackObama might indicate that there were more @-replies in the dataset, but our analysis indicates otherwise.

Table 1
Syntactical Feature Frequency

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Tweets</th>
<th>Median Position</th>
<th>Link %</th>
<th>Hashtag %</th>
<th>At-reply %</th>
<th>Retweet %</th>
</tr>
</thead>
<tbody>
<tr>
<td>BarackObama</td>
<td>3,935,897</td>
<td>10.24%</td>
<td>36.36%</td>
<td>30.70%</td>
<td>24.13%</td>
<td>59.29%</td>
</tr>
<tr>
<td>MittRomney</td>
<td>1,311,983</td>
<td>21.77%</td>
<td>29.94%</td>
<td>39.12%</td>
<td>29.42%</td>
<td>47.82%</td>
</tr>
</tbody>
</table>

Longitudinal Analysis

There is a definite difference in the position of the @-mentions of the two candidates in both dataset over a five-month period indicating different uses. In order to more fully understand the longitudinal trend of this median position, we plotted the median position of the two handles on a daily basis over the examined time period. Figure 1 represents the longitudinal position of @BarackObama and Figure 2 represents the longitudinal position @MittRomney. For the purpose of illustration and comparative analysis we plotted a standard regression line of best fit (red) and LOESS derived line (blue) that more closely identifies the localized movement of the medians over time.
The longitudinal analysis illustrates that although the daily median position of the @-mentions of the two candidates varies, it does not vary significantly as the standard regression line for Barack Obama only has a slight positive slope and the standard regression line for @MittRomney has almost no change longitudinally. The blue line represents a LOESS regression, which identifies more localized shifts in the position of the hashtag. Although the blue line oscillates for both @BarackObama and @MittRomney, the
position seems to have an equilibrium that is marked by the standard regression line. That is, we identify that although there are daily shifts, the general position of the @-mention does not differ significantly over time.

**Identifying the Shift**

We need an explanation for why @barackobama and @mittromney maintain their position over time at the first position and early-to-middle position, respectively. We hypothesized that this was the result of more individuals directly addressing @BarackObama, but the distribution of @-replies indicates a similar percentage of tweets that had these characteristics.

We identify that in the @BarackObama dataset, 53.05% of all of the retweets were retweets of @BarackObama tweets. In contrast, in the @MittRomney dataset, we see that only 16.97% where retweets of @MittRomney’s tweets. Since the most common nomenclature of a retweet is “RT @[username]” and is commonly done at the beginning of the tweet (Kooti et al., 2012). This high percentage of @BarackObama being retweeted would indicate that this is likely the cause of @BarackObama tending to appear earlier in the tweet but not quite first. It’s also why @Mittromney appears later; tweets are more often about him than retweets of him or tweets directed at him.

**Discussion**

Our findings represent an exploration of the position of Twitter handles in one specific context and contribute to an emerging discourse about the syntactical features of Twitter. First, we identify that the position of handles of different users does vary and that longitudinally this variance tends to have an equilibrium state even though there is some oscillation. Second, we identify that even though the syntactical feature distribution in datasets may not vary significantly, that further analysis identifies significant differences can be found in the content of different syntactical features.

We determine that @BarackObama is, on average, mentioned earlier in tweets because he is more often being retweeted. Even though the @MittRomney dataset had a significant amount of retweets we see that these retweets were mostly about him and not from him. This illustrates a distinctly different level of discourse that needs to be accounted for in analysis of big datasets.

**References**


Reading Practices and Intellectual Freedom Research

Emily J.M. Knox
University of Illinois at Urbana-Champaign
knox@illinois.edu

Abstract

Based on discourse from 13 challenge cases from 2007-2011, the study argues that studying contemporary reading practices, especially interpretive strategies, is vital to understanding why people attempt to censor books.

Keywords: reading practices, intellectual freedom, censorship

When the Graduate Library School (GLS) at the University of Chicago was established in 1929, the faculty considered reading research to be an integral component of graduate education in library and information science. Under the guidance of Douglas Waples and following the program of the sociology department at the university, several reading studies were conducted in the Chicago area. In a 1999 article, Wayne Wiegand notes that this research ended in 1932 under the auspices of a new dean, Louis Round Wilson. Wiegand states—in no uncertain terms—that GLS and librarianship “walked away from an opportunity” (Wiegand, 1999, p. 10). This study, by drawing on concepts from the interdisciplinary field of book history, demonstrates how our understanding of reading practices can impact research on intellectual freedom and censorship.

Reading is a practice that is somewhat difficult to theorize. Robert Darnton notes that there are no direct routes to comprehending reading—only approaches that can be used by researchers to gain a better understanding of the phenomenon. Darnton’s first approach investigates how people learn to read. Researchers can also study reading through autobiographies. This approach explores what people say they read. Another approach to reading research is through literary theory, particularly reader-response theory. A fourth method explores analytical biography. Finally, researchers can examine the “ideas and assumptions underlying reading in the past” (Darnton, 1991, p. 171). For example, attempts to censor reading materials demonstrates not only ideas with which people disagree but also what people believe constitutes “good reading.”

In her article on textual interpretation, Elizabeth Long (1992) demonstrates the social and collective nature of the practice of reading. Even though reading is often seen as a solitary activity (a concept the author vividly illustrates through a series of paintings that show lone readers), Long establishes the collective nature of reading by demonstrating its reliance on both social infrastructure and social framing. By social infrastructure, Long means that reading is an activity that is learned through social relationships and relies on the social base of literary culture. Social framing constructs certain materials as being “worth reading” and is socially defined. It is possible that this concept of “worthy” reading can be linked to the idea of “appropriate” reading that is so prominent in the discourse of censors.

One method that might be used to study reading is to focus on the interpretive strategies that people employ when encountering written texts. Stanley Fish notes that “interpretive strategies are not put into execution after reading...they are the shape of reading and because they are the shape of reading they give texts their shape, making them rather than, as it is usually assumed, arising from them” (Fish, 1982, p. 168). Interpretive strategies are defined here as a set of decisions that one makes both before and while one is reading. These decisions have many different influences including social constructions of written texts and the perceived authority of the book.
This study discusses reading as a social practice that has changed over time and encompasses different physical modalities and interpretive strategies. In order to understand why people challenge books, it explores differing understandings of how reading works, "what it means" to read a text, and especially how one constructs the idea of "appropriate" reading materials. Data for the study consists of discourse from 13 challenge cases to books in American public libraries and schools that took place between 2007 and 2011. Three sources of discourse from the cases were used in the study. The first consisted of documents, obtained via Freedom of Information Act requests to governing bodies, produced in the course of challenge cases. Recordings of book challenge public hearings constituted the second source of data. Finally, the third source of data was interviews with challengers.

The study demonstrates that challenge cases are particularly influenced by what might be called a "common sense" orientation to text wherein there is little room for polysemic interpretation. The meaning of texts is always clear and there is only one avenue for interpretation. This common sense interpretive strategy is coupled with what Cathy Davidson (2004) as “undisciplined imagination” wherein the reader is unable to maintain distance between the events in a text and his or her own response. These reading practices broaden our understanding of why people attempt to censor books in public institutions. The study posits that censorship behaviors, such as challenging or marking through books, are intimately tied to the how one understands the practice of reading and its effects on character development.

References


Open Access Publishing Fees: Responses, Strategies and Emerging Best Practices

Laura Waugh
University of North Texas
laura.waugh@unt.edu

Abstract
The Open Access (OA) movement has led to a rethinking and restructuring of traditional publishing funding models. A growing number of OA journals require authors to pay an Article Processing Charge (APC) in order to have their articles published in their journal. In addition, hybrid journals (i.e. traditional, subscription-based journals), are beginning to offer the option to make an article OA if authors pay an APC. Building on the research initiatives of the Scholarly Publishing and Academic Resources Coalition (SPARC), the University of North Texas (UNT) Libraries conducted a review of thirty North American universities’ OA fund initiatives, fifteen OA journal funding models, and twelve hybrid journal funding models in order to better understand this trend. This poster illustrates the findings of this research and identifies emerging best practices among universities that have implemented an OA fund.

Keywords: open access, funding models, OA journals, hybrid journals, article processing charges

Purpose
The Open Access (OA) movement has led to a rethinking and restructuring of traditional publishing funding models. A growing number of OA journals require authors to pay an Article Processing Charge (APC) in order to have their articles published in their journal. In addition, hybrid journals (i.e. traditional, subscription-based journals), are beginning to offer the option to make an article OA if authors pay an APC. This practice of charging an additional APC in order to provide open accessibility to articles is seen by many publishers as a transitional method from subscription-based models to more inclusive models of funding to incorporate OA initiatives. In response to the increasing number of journals charging APCs that authors are responsible for paying in order to have open accessibility to their work, a growing number of universities are creating OA funds in order to help cover a portion of the costs.

Methods
Building on the research initiatives of the Scholarly Publishing and Academic Resources Coalition (SPARC), the University of North Texas (UNT) Libraries conducted a review of thirty North American universities’ OA fund initiatives, fifteen OA journal funding models, and twelve hybrid journal funding models in order to better understand this trend. The review included consulting research by SPARC (Open-access funds in action, 2012), the University of California Berkeley (Selective list of open access, 2010), BioMed Central (Comparison of BioMed Central’s, 2010), Virginia Tech University Libraries (Open access subvention fund, 2012), Tufts University (Provost’s Open Access, 2010), and the University of Connecticut Health Center (Open Access Author, 2012). The UNT Libraries’ research focuses on the funding sponsors of the OA funds, the eligibility requirements of authors and articles, the reimbursement criteria, examples of OA journals and APCs, and any stipulations placed on the OA fund for authors.

Results
Of the thirty universities reviewed, twenty-seven are sponsored completely, or in part, by their university’s library. Fourteen of the thirty universities reviewed receive co-sponsorship through other administrative or institutional initiatives.

Copyright is held by the author.
All thirty universities reviewed accept submissions to their OA reimbursement fund only from authors affiliated with their university. Of the thirty universities reviewed, all accept submissions from faculty, twenty-six accept submissions from students, nineteen accept submissions from researchers, and fifteen accept submissions from staff.

Of the thirty universities reviewed, all support OA journals. Fourteen also support hybrid journals, in some capacity. Of the fourteen universities reviewed that also include hybrid journals in their OA reimbursement fund, nine include specific stipulations that must also be met. The stipulations include: 1) no embargo period on the article, 2) that the author retains distribution rights or copyright, and 3) that the publisher agrees to reduce the subscription cost to the university.

All thirty universities reviewed stipulate that articles considered for their OA reimbursement fund be published in peer-reviewed journals. Sixteen of the universities stipulate that the journal be listed in the Directory of Open Access Journals (DOAJ) and fourteen of those also require additional stipulations for the journal including that the publisher, 1) be a member of the Open Access Scholarly Publishers Association (OASPA) or adhere to their Code of Conduct, 2) have a publicly available standard fee schedule, and 3) have a policy to waive fees in the case of economic hardship.

Of the fifteen open access journal publishers reviewed, the average APC is approximately $2000. The APCs range from $695 to $5000. Of the twelve hybrid journal publishers reviewed, the average APC is approximately $2600. Hybrid journal APCs range from $645 to $5000.

Twenty of the thirty universities reviewed include a specific cap on how much funding one author may receive in a given period of time, or how many times one author may apply for reimbursement. Twenty-one of the thirty universities reviewed include a specific cap on the funding awarded to each article.

Four of the thirty universities reviewed require authors who are awarded an OA reimbursement fund to also include their article in the university’s institutional repository. One university also stipulates that the university be listed as the primary affiliation of the author in order for the article to be eligible for the OA reimbursement fund.

All thirty of the universities reviewed specify that the OA reimbursement funds will be disbursed on a first-come first-serve basis, and they offer a submission form for authors to fill out either electronically or in a downloadable format. In addition, all thirty universities have established a web presence for their reimbursement funds with information on the policies and criteria.

Conclusion

This poster illustrates the findings of this research and identifies emerging best practices among universities that have implemented an OA fund. As publishing practices, funding models, and scholarly communications continue to evolve, universities must find innovative ways of keeping up. OA funds are one approach that universities can use to promote open accessibility to the valuable research outputs of their faculty and students. Although trends are apparent in the research data, best practices will not be fully developed until more universities adopt this approach.

References

Comparison of BioMed Central’s article processing charges with those of other publishers. (2012). Retrieved from: http://biomedcentral.com/about/apccomparison/


Selective list of open access and paid access fees. (2010). Retrieved from University of California Berkeley: http://lib.berkeley.edu/scholarlycommunication/oa_fees.html
Information Behavior of Rural Dementia Caregivers Over the Course of Counseling Interventions

Michelle M. Kazmer  
Florida State University  
mkazmer@fsu.edu

Robert L. Glueckauf  
Florida State University  
robert.glueckauf@med.fsu.edu

Jinxuan Ma  
Florida State University  
jm07@my.fsu.edu

Gabriel Schettini  
Florida State University  
gs04@my.fsu.edu

Michelle Silva  
Florida State University  
ms08r@my.fsu.edu

Abstract

This poster reports findings from an analysis of 81 qualitative interviews with rural dementia caregivers who participated in two research projects that evaluated cognitive behavioral therapy (CBT) problem-solving interventions designed to help them manage their caregiving and reduce depression. The research questions posed in this analysis are: (RQ1) How do rural dementia caregivers who have participated in a CBT program interact with the information around them and available to them? (RQ2) How do these caregivers manage, explicitly or tacitly, their information needs and uses? The analysis demonstrates similarities and differences in information behaviors and information environments across the caregivers; for example that physical information artifacts are important to different caregivers for varied reasons, and that religious communities fulfill multiple information-related roles. The findings demonstrate the importance of accounting for the unique constraints on older caregivers in designing interventions to improve and support their caregiving experiences. Creating links among people, such as connecting caregivers to similar others facing similar challenges, and providing them with a structure to allow them to focus on their goals, appear critical to the success of such interventions.

Keywords: dementia caregivers, information use environments, health information behavior, qualitative health research

Introduction

Most adults with dementia (80%) receive care from family members at home (Alzheimer’s Association, 2011). These caregivers (CGs) often experience significant psychological distress, sleep fragmentation, reductions in social activities, and disrupted family relationships, leaving them at substantial risk for depressive disorders (Gallagher-Thompson et al., 2000; McCurry, Logsdon, Teri, & Vitiello, 2007) and compromised physical health (e.g., Kiecolt-Glaser, Glaser, Gravenstein, Malarkey, & Sheridan, 1996; Vitaliano, Schulz, Kiecolt-Glaser, & Grant, 1997). Increased understanding of how CGs use information resources to manage their caregiving demands, and how context shapes their information use, can improve the design of CG-related information sources and services.

This study seeks to further this understanding by examining rural dementia caregivers who participated in two research projects that offered cognitive behavioral therapy (CBT) problem-solving interventions to help them manage their caregiving and reduce depression: 1) a study of faith community nurses (FCN) trained to conduct cognitive behavioral and spiritual counseling for rural dementia caregivers; and 2) a study called “African-American Alzheimer’s Caregiver Training and Support” (ACTS), which provided skills training and support, by phone or in person, to African-American dementia caregivers with depression (Glueckauf et al., 2012).

Acknowledgements: This work was sponsored by the U.S. National Institute of Mental Health (R34MH078999), the Florida State University College of Medicine, and the University of South Florida Health Byrd Alzheimer’s Institute.


Copyright is held by the authors.
Background

Most information-oriented examinations of CGs focus on information needs (e.g., Conley & Burman, 2011; Hirakawa, Kuzuya, Enoki, & Uemura, 2011; Thompsett & Lovestone, 2002; Wackerbarth & Johnson, 2002; Wald, Fahy, Walker, & Livingston, 2003). Some focus on information materials (e.g., van der Steen et al., 2011), while others focus on information providers or intermediaries (e.g., Abrahamson, Fisher, Turner, Durrance, & Turner, 2008). Another subset of literature indicates the importance of examining and supporting information use among the elderly, but without a focus on CGs (Niemelä, Huotari, & Kortelainen, 2012; Xie & Bugg, 2009). Information needs are part of a broader approach that studies information behavior, which includes other phenomena such as information seeking, retrieval, motivations, avoidance, incidental acquisition, serendipity, overload, evaluation, and management (see Dervin, 2005; Ginman, 2000).

Understanding information behavior requires that we "account for... the social context within which information is generated, sought for, acquired, evaluated, organized, disseminated, and used" (Rosenbaum, 1996, p. 152). Many frameworks have been developed to consider such context, including information fields (Johnson, 1997), information grounds (Pettigrew, 1999), information ecologies (Nardi & O'Day, 1999), information worlds (Burnett & Jaeger, 2008), and information use environments (Taylor, 1986, 1991). Such frameworks have been evaluated for their applicability to CGs (e.g., Harland & Bath, 2008; Ormandy, 2010). When people seek information about health (and other) activities and events, this context includes many factors such as individuals' roles (Wicks, 1999), culture and ethnicity (Jeong, 2004; Yi, Stvilia, & Mon, 2012), age (Asla, Williamson, & Mills, 2006), living environment (Mooko, 2005; Shin, 2009) and – for health-related information behavior – the type of medical condition and associated symptoms (Johnson, Andrews, & Allard, 2001).

As part of an overall research agenda focused on the information behaviors of dementia caregivers, this poster addresses two specific research questions: (RQ1) How do rural dementia caregivers who have participated in a cognitive behavioral therapy program interact with the information around them and available to them? and (RQ2) How do these caregivers manage, explicitly or tacitly, their information needs and uses?

Methods

The FCN and ACTS projects were designed as quantitative trials of the CBT interventions, but each included a qualitative portion. Qualitative data were collected via interviews with participants, totaling 81 interviews across both projects. All interviews were transcribed. The preliminary "open" coding – coding that emerges from the content of the interviews rather than from a pre-existing code list (Charmaz, 2006; Strauss & Corbin, 1998) – suggested that further qualitative analysis of these data, directed toward information behaviors, was needed. All 81 interview transcripts were subsequently coded with a focus on expanding and refining the existing open codes that reflect information behaviors, and adding codes suggested by sensitizing concepts (Strauss & Corbin, 1998, pp. 48-52) such as information use, information exchange, information avoidance, and so forth. NVIVO8 qualitative analysis software was used as a platform to support this analytic work.

Findings

The findings to be presented in this poster represent the primary similarities and differences across the CGs in both projects. All the CGs engaged in problem identification and goal setting. They worked with their counselors to record the problems they faced and to develop goals they could achieve to help ameliorate the effects of those problems. Their "selves" became primary information sources because they were responsible for identifying their problems and for creating a physical information artifact to record those problems and associated goals. The CGs learned to turn to themselves for verification of their progress: Is what I am doing helping me achieve my goals? Have I completed the tasks I need to complete to help me achieve my goals? Now that I have completed my goals, for what (new) problems do I need to set new goals?
Common among CGs was the role of the counselor as an information intermediary facilitating open information exchange among counselors and CGs. The counselors interpreted printed instructions or rote problem-solving steps, making them understandable and actionable. CGs relied on the counselors for accurate and timely information in various areas including legal, mental health, physical health, lifestyle, and so forth.

CGs’ religious communities and spiritual beliefs and practices also shaped their information environments. ACTS and FCN CGs involved in religious communities rely on them for information and social support. Among the large majority of CGs in both studies who held strong religious beliefs (primarily Christian), those beliefs influenced their information behaviors such as their willingness to seek information, tendency to believe or disbelieve certain types of information, and how they dealt with healthcare professionals as people who had information and were empowered to act on it (see also Lustria et al., 2010).

Major differences in information behavior are driven by intrinsic differences in the two projects. ACTS includes a guidebook for CG participants that takes on high importance in CGs’ lives: they use it as a reference, as a source of authority in dealing with family members or healthcare providers, and as a physical reminder of their goals. The FCN study included no such guidebook; occasionally the FCNs gave printed materials to their CGs, but it was not an intrinsic part of the intervention.

Another difference was that the ACTS project included group and individual counseling sessions while FCN only included individual sessions. The groups of ACTS CGs were an important source of information to each other; they exchanged alternative problem solving techniques and shared information among CGs whose care recipients were at different stages of dementia, which the CGs indicated was important in reducing their anxiety and fear. The CGs in FCN often sought other groups of CGs (such as church groups) to provide this type of support. When they did so, they were outside the structured problem-solving environment offered by the CBT intervention and some groups devolved into shared suffering rather than focused problem solving.

Conclusion

The findings in this project remind us that it is important to focus on CGs’ information behaviors in context and identify their information needs in light of their problems and goals. The CGs information behaviors are shaped by economic, cultural, and social factors including religion, cultural factors associated with race and region, access to the internet and other information sources. When considering an activity or intervention, it is important to consider the constraints on the target users and do what will work best for them rather than what seems interesting in terms of technology, or rich in terms of interaction (see Forducey, et al., 2012). Including the right people (e.g., similar others facing similar challenges) and providing the CGs with a structure to allow them to focus on their goals appear critical to the success of such interventions.

References


Can’t See the Forest for the Trees?  
A Citation Recommendation System

Cornelia Caragea  
Computer Science and Engineering  
University of North Texas  
ccaragea@unt.edu

Adrian Silvescu  
Naviance Inc.  
silvescu@gmail.com

Prasenjit Mitra  
Information Science and Technology  
Pennsylvania State University  
pmitra@ist.psu.edu

C. Lee Giles  
Information Science and Technology  
Pennsylvania State University  
giles@ist.psu.edu

Keywords: citation recommendation, singular value decomposition, collaborative filtering

Background and Motivation

As science advances, scientists around the world continue to produce large numbers of research articles, which provide the technological basis for worldwide collection, sharing, and dissemination of scientific discoveries. When writing a paper, an author searches for the most relevant citations that started or were the foundation of a particular topic, which would very likely explain the thinking or algorithms that are employed. The search is usually done using specific keywords submitted to literature search engines such as Google Scholar [3] and CiteSeer [2]. However, text-based search engines return poor results when there is vocabulary mismatch between a query and the relevant documents. Moreover, finding relevant citations is distinctive from retrieving articles that are only topically similar to an author’s proposal. For example, Teufel et al. [6] showed that citations can be of various types, and provided an annotation scheme for the citation function, which consists of twelve different categories. Among these categories, some citations are topically similar, others are used as survey articles to provide background information to the reader, while yet others contain tools/algorithms/data that are adapted or modified in the new proposal [6].

What is a good strategy to uncover both topically-related and, at the same time, distant, but highly-relevant citations for a particular query, while filtering out irrelevant information, given today’s very large collections of published articles? One promising line of research to information filtering is the design of recommender systems. Collaborative Filtering (CF), the commonly used method for recommender systems, relies on the assumption that similar users express similar interests on similar items. Two types of memory-based CF are: user-based CF, which computes the similarity between users, based on user profiles or history, and item-based CF, which computes the similarity between items, based on various item information.

However, memory-based CF algorithms have several limitations such as data sparsity and scalability [4]. Consider, for example, the problem of citation recommendation in CiteSeer [2], where the underlying citation graph tends to be noisy (e.g., due to errors in citing, and missing citations, etc.) and sparse (i.e., most papers in CiteSeer have on average about 30 citations, which is far below 1% of the more than two million papers indexed into the database). Memory-based CF algorithms, which are nearest neighbor based algorithms, rely on exact matches while computing (item or user) vector similarities, and hence, could result in loss of citation recommender system coverage and accuracy [5]. Moreover, the nearest neighbor algorithms require similarity computation that grows with the number of research papers. With more than two million papers available in CiteSeer, CF-based citation recommendation systems suffer from scalability problems.


Copyright is held by the authors.
Against this background, we address the problem of citation recommendation using singular value decomposition (SVD) [1] on the adjacency matrix associated with the citation graph to construct a latent “semantic” space, where citing and cited papers that are highly correlated are placed closed to each other. The main idea behind SVD is to project the original high-dimensional data into a lower-dimensional space, in which patterns in the data can be more easily identified.

We exploit information available in the online CiteSeer digital library of scientific publication to train and evaluate our models. The assumption is that, when writing a paper, an author has some background knowledge about the topic he writes about and that an initial set of citations (i.e., a “basket” of citations) is provided as input to the recommender system. We require the system to retrieve other relevant works that the author might have missed (works that should be cited or the author should be aware of).

**Experiments and Results**

The CiteSeer Dataset: The citation recommendation data set used in our experiments is compiled from the CiteSeer citation graph and the metadata available for each paper indexed in CiteSeer [2], as of December 2011. There are 1,345,249 unique citing papers and 9,150,279 unique citations in the CiteSeer citegraph. The total number of links in the graph, i.e., [citing paper → citation], is 25,526,384. From the CiteSeer citegraph and the available metadata, we constructed a smaller data set for the task of citation recommendation as follows: we filtered out papers that do not have title and abstract, as well as papers that are cited by other papers in the corpus less than 10 times and more than 100 times. In addition, we filtered out papers that cite less than 15 or more than 50 other papers. In our resulting citegraph, there are 81,508 unique cited papers (citations), 16,394 unique citing papers, and 341,191 links.

Experimental Design: We address the following question: How does SVD compare with collaborative filtering approaches on the citation recommendation task? We split the data set into training and test sets by randomly selecting one non-zero entry from each citing paper, to be part of the test set, whereas the remaining non-zero entries are considered part of the training set (i.e., the “basket items” for each citing paper). We sampled a validation set, used to estimate model parameters, in a similar manner.

Figure 1 shows the results of the comparison of SVD with CF algorithms, i.e., CF User-based Simple Weighted Sum (SWS), CF User-based most-frequent-item (F), and CF Item-based. Each algorithm returns a list of top N recommendations for each citing paper, with N ranging from 5 to 25 in steps of 5. If the hidden citation (in the test set) is part of the top N recommendation list returned by an algorithm, the algorithm was considered accurate for the particular citing paper. We repeated each experiment 5 times to ensure the results were not sensitive to a particular train-test split and averaged the results across the five runs. As can be seen in the figure, SVD outperforms the CF models in terms of both Recall and Precision, for all values of N. In future, because SVD-like models easily allow incorporation of additional information, we plan to integrate other types of information (e.g., textual information) into our models.

![Figure 1: Comparison of SVD with collaborative filtering algorithms in terms of Recall and Precision.](image-url)
References


“I have to read the signs like you reads a book”: Examining Experiential Information through Non-fiction Accounts of Guided Trips into the Interior of Algonquin Provincial Park Ontario

Angela Pollak
Western University, London, Ontario
University of Western Ontario
apollak@uwo.ca

Abstract

Skill-based careers and remote rural settings are often overlooked in LIS research. Analyzing the content of two historical non-fiction accounts of extended canoe trips, this poster explores the information life world of old-time wilderness guides in Algonquin Provincial Park in Ontario, Canada. Analysis provides a glimpse into the unique ways they have of acquiring and sharing their particular brand of knowledge. Results suggest their life experience is rich with information behaviors and skills both similar to, and unique from, other contexts. They also support the advancement of information theory by further challenging the boundaries between formal and informal information systems and by conceptualizing information in terms of intangible, undocumented experiences.

Keywords: information behavior, life world, experiential information, wilderness guides, rural settings, content analysis

Introduction

For more than a hundred years, remote and rural Algonquin Provincial Park, Ontario has played host to an annual migration of people who wish to enjoy the great outdoors. In 2011 alone, over 245 000 visitors, mostly from Ontario, camped in the interior of the Park (Ontario Parks 2011). Although it is becoming easier to undertake such trips unaccompanied due to advances in technology, many visitors over the years have engaged local help. For a fee, the local residents have been asked to do what they do best: use information gained through experience to guide visitors safely through vast stretches of wilderness. This experiential information has typically been acquired over time, by hand, either in trial-by-error fashion or by spending time in the company of others who had previously learned the skills. While experiential information is often described in both fiction and non-fiction accounts of different wilderness expeditions, it has not yet been the focus of attention in LIS research literature. This poster explores the information world of two historical figures who guided in Algonquin Provincial Park during the 20th century.

The Information Problem

“Guiding” is the act of leading visitors on trips through the bush for pleasurable leisure purposes, generally by canoe and on foot, over a period of time ranging from one day to several weeks. Guiding involves activities such as planning routes and itineraries, preparing food and supplies, and securing
permits and transportation. It requires an intimate knowledge of the terrain, habitats, and weather patterns, as well as practical 'how-to' information like how to set up shelter and how to catch fish. It also often involves sharing information about the flora, fauna and folklore of the area with the visitors. While it has been a licensed occupation in Algonquin Park since the early 1900s, there is no school one can attend to learn how to guide. And so, while the occupation is regulated, the training is not credentialled in any way. It has historically been a hands-on skill acquired primarily through everyday life interactions and both private and social experiences.

From an information perspective, the guiding context presents a fascinating and difficult 'life world' (Savolainen 2008) from which to study information behaviors because of both the location and manner in which guides acquire their knowledge, and the fact that the researcher must have enough interest and physical stamina to participate in the voyageur experience itself in order to observe this population. Given these challenges, and the fact that the guide’s workplace occupies over 7600 square kilometers (Ontario Parks 2011) of dense Canadian Shield, it is not surprising that this context has not previously been studied in LIS literature.

Nevertheless, there are accounts in the world of literature that provide a glimpse of the gap in our understanding. The following excerpt written by Bernard Wicksteed, an RAF pilot returning to civilian life after serving in WWII, describes the difference he found between his own credentialed background, and the experiential background of his guide (Joe Lavally) during a trip into Algonquin Park in 1945:

"Joe shrugged his shoulders. What could you do with a man who couldn’t read the book of nature when it was spread out in front of him? Who saw ducks take off in alarm and didn’t ask himself what had frightened them? Who watched frogs jump from a ten-foot rock and didn’t look for the cause? He knew too much about city folks to think you could ever teach them to use their eyes and their heads. The best you could do was try to please” (Wicksteed 1948).

Studying this information gap is particularly interesting at this time in view of the emergence of a new breed of guide accustomed to a wider array of tools and technology. As the old-time guides who worked without this benefit retire and expire, we lose the opportunity to understand their ‘old’ ways of guiding. Thus, the characteristics of experiential information in this context, and the myriad ways of ‘knowing’ it, are at risk of being lost to us forever.

Methodology and Theory

To study the phenomenon of experiential information among old-time wilderness guides, I examine two first-person, historical non-fiction accounts of guided trips into Algonquin Park centered around the geographic region of Whitney, Ontario (Wicksteed 1948, Corbett 2001). These accounts are written by visitors and contain frequent remarks about the information skills they witnessed during their extended excursions.

I justify using these monographs for several reasons. First of all, they represent historical accounts of guiding, meaning that the timeframe during which the two protagonists were learning about guiding and engaged in the activity for pay was a period in the Park’s earlier years (early to mid 20th century). This is significant because it was a time period when the Park was measurably less accessible to the masses. Further, tools used by the guides were often hand-crafted during this time period, and there were very few amenities available. Technical communication options (GPS, cell service, satellite service) did not exist. Consequently, guides of that time period developed unique information skills in response to the environment in which they worked. This poster describes the role of information in the historical act of guiding, which I believe is significantly different than the guiding experience of today.

Secondly, there are very few people alive today who earn a living in this geographic area guiding in the traditional sense. Attrition and urban migration are diminishing the number of local young people who choose guiding as a profession, thus reducing the number of guides available who learned how to guide as part of their everyday life experience. Consequently, there are very few sources from which to obtain primary accounts of the guiding experience. In the absence of primary sources and without evidence questioning the truthfulness of the monographs I’ve chosen, the books are essentially part of a limited number of historical records available to us about this information activity, and they are the only first-person accounts. During the course of my doctoral research project involving interviews with more
than twenty local residents, several interviewees recommended these books as a way to better understand the local residents’ way of life. They are considered authoritative even among the local population.

Perhaps the most compelling reason to use these data sources is that the information is observed and presented by outsiders to the community. These people are in a unique position to observe, identify and comment on behavioral idiosyncrasies that may never come to light during interviews with primary subjects, either because the primary source doesn’t recognize their significance, or because of self-censoring.

I examine these texts through content analysis focusing on both the characteristics of what the narrators tell us the guides know, and also suggestions about how they come to know it. Data are then analyzed according to grounded theory using the constant comparative method, whereby “incidents that are found to be conceptually similar” are given “the same conceptual label” until each emerging theme is fully elaborated (Corbin and Strauss 2008). Themes are analyzed primarily from the theoretical perspective of Elfreda Chatman's studies into life in the round (Chatman 1999) and Reijo Savolainen’s studies of everyday life information seeking contexts (Savolainen 2008, Savolainen 2009).

Conclusions

For the guides of Algonquin Park, the information gained from their life world is natural, commonplace, and often socially derived (Savolainen 2008) from the community in which they live. It is also largely absent from LIS literature, which has focused primarily on information behaviors of urban, often professional, populations. This poster outlines how experience-based information behaviors and skills found in this particular life world are both similar to and different from those described in the literature. Behavior similarities include evidence of secrecy and deception as described by Chatman (Chatman 1999), and the concept of way of life as described by Savolainen (Savolainen 1995). It also describes unique information skills including the ability to observe and interpret environmental signals, store and recall vast amounts of information about diverse wildlife species, as well as foraging and caching skills.

For information behavior theorists, a glimpse into this life world presents a valuable opportunity to improve our understanding of a previously unstudied, less accessible population and life world, as well as the unique ways they have of acquiring and sharing their particular brand of knowledge, most of which occurs without documentary support. It also supports the advancement of information theory by further challenging the boundaries between formal and informal information systems and by conceptualizing information in terms of intangible, undocumented experiences.

References

Enhancing Content Visibility in Institutional Repositories: Overview of Factors that Affect Digital Resources Discoverability

Ahmet Meti Tmava  
University of North Texas, College of Information  
Ahmet.Tmava@unt.edu

Daniel Gelaw Alemneh  
University of North Texas, Digital Libraries Services  
Daniel.Alemneh@unt.edu

Abstract

In the last decade, a growing body of the scholarly information and instructional materials produced by universities existed primarily in digital format. New digital technologies increased the productivity of scholars. The purpose of an Institutional repository (IR) is to manage their scholarly work in ways that facilitate interdisciplinary collaboration, and accelerate the pace of discovery and innovation. Academic institutions have increasingly recognized that IRs are a vital part of the scholarly dissemination infrastructure. The goal of an IR is essentially to collect, preserve, and make persistently accessible a variety of scholarly materials. This paper explores digital curation activities that enhance the visibility of IR in an ever-changing digital landscape.

Keywords: institutional repositories, metadata standards, metadata consistency, search engines

Introduction

Institutional repositories (IR) are defined by Piorun et al. (2007) as electronic systems that capture, preserve, and provide access to the intellectual output of a community in a digital format. The goal of an IR is to collect, preserve, and make persistently accessible a variety of scholarly materials (Palmer et al., 2008). Open access digital repositories have implemented the Open Archive Initiative-Protocol for Metadata Harvesting (OAI-PMH) as a mechanism to achieve interoperability in the exchange of meta-information with other systems (Muhammad et al, 2007). With all the different software platforms employed by IRs their systems need to be able to support interoperability in order to provide access via multiple online search engines.

IRs are responsible for not only creating metadata but also for ensuring that metadata is available in standard schemas and formats that comply with the OAI-PMH protocols. With all the digital content available online today, it is a constant struggle to attract users. One of the challenges facing IRs today is resistance by faculty to voluntarily depositing their works (Albanese, 2009). In order to make their cases, IRs need to assure faculty that their work can be widely disseminated. In this paper we explore factors that enhance the visibility of IR items in an ever-changing online environment.

Institutional Repositories and the Changing Landscape of Scholarly Communication

New digital technologies have increased the productivity of scholars and enabled them to manage their own digital content in ways that facilitate interdisciplinary collaboration and accelerate the pace of discovery and innovation. IRs provide long-term sustainable storage, preservation, and open access to resources. Foster & Gibbson (2005) found that faculty want to be able to make their own work available to others and have easy access to other people’s work. IRs also serve as tangible indicators of an institution’s productivity, and thereby increasing an institution’s visibility, prestige and value.


Copyright is held by the authors.
In spite of the increase popularity of IRs worldwide (see Figure-1), a survey conducted at universities in United States and Canada in 2009 found that only 56.7% of faculty were aware that their institution had a digital repository where they can publish their work. And out of which, less than 10% of faculty have ever submitted material to their institution’s digital repository (Primary Research Group, 2009, p.17).

In this regard, one of the main challenges for Institutional Repository managers is to make faculty aware that IRs offer them new dissemination opportunities, to reach a much broader audience than what is available through other formal means of publishing. Through open access IRs make their digital object available to a worldwide audience. This can increase citations made to their scholarly work, thus it would maximize their impact factor. The “impact factor” refers to the number of subsequent citations a work receives (Donovan and Watson, 2011).

IRs reform scholarly communication by stimulating innovation in a disaggregated publishing structure (Crow, 2002). In light of the continually evolving information environment and user needs, digital curators (among other stakeholders that are actively involved in supporting IR development) must promote IR use, maintain consistency and participate in the creation of and adherence to national/international standards and institutional-specific policies. Furthermore, it is vital that they understand open standards for searching and effective retrieval of digital content. Open standards enhance interoperability between different IR software platforms, such as DSpace (see Figure-2), and online search engines, such as Google, to index metadata elements consistently and offer effective search capabilities.
Figure 2: Worldwide Usage of Open Access Repository Software*.
*Source: Open DOAR (Directory of Open Access Repositories), December 2012.

Current Best Practices and Ways to Increase Content Visibility

Search engines are the most popular way users search for information; they account for 88 percent of users’ search time when they are looking for information (Bifet and Castillo, 2005). Most users like to search by keyword, and they expect to retrieve useful results. Google is the most popular search engines because it is very easy to use, it is free, it is extremely fast and produces useful results. For the most part Google generates search results by matching search terms entered with Web page content, usually referred to as full text searching. This form of searching usually has shortcomings, mainly in precision. Precision refers to the proportion of the relevant documents retrieved in a search to the total number of documents retrieved in a search.

Contrary to the long-held belief that Google ignores the data encoded in a Web page’s meta tags, Zhang and Dimitroff (2005) show that “metadata is a good mechanism to improve webpage visibility.” Metadata describes the nature of the digital items stored in a repository, including content, structure, and access rights. In other words, metadata should provide multiple access points (e.g. author, title, subject). To fulfill its purpose it is important for the digital curators to understand the ability of search engines to index metadata elements and retrieve digital content using their embedded metadata elements.

Web crawlers operated by search engine companies harvest metadata and other information about online objects and send that information back to the search engine. Specialized algorithms (that take many factors into account) analyze the harvested data and decide whether or not to add the metadata to the search engine’s index. Farajpahlou and Tabatabai (2011) note that XML, as a syntax ground for implementing the metadata elements of DC and MARC 21, maximizes the interoperability between search engines and metadata initiatives. According to Taheri and Hariri (2012), Google and Yahoo treat XML-based metadata with language-based tags (like DCXML) and without language-based tags (such as MARCXML) in the same way. The same authors noted that all metadata elements of the MARCXML and DCXML standards are compatible with the Google and Yahoo indexing software.

Institutional Repository systems need to be able to support interoperability in order to provide access via multiple search engines. According to Hirwade (2011), “interoperability is the ability of systems, services and organizations to work together and exchange information and use exchanged information without special effort of either system”(p.59). Metadata interoperability facilitates the exchange of information between repositories and enables World Wide Web searching. To make items discoverable in the diverse online environment, the database needs to provide standards-compliant database servers and expose the fullness of that metadata to a metasearch search client (Dorman, 2008). Metasearching, or the process of metadata being searched, enables connections to multiple
resources. However, as described by Dorman (2008), in cases when the IR does not supply a standards-compliant database server there are two alternative access methods that can be used: a proprietary gateway or HTML parsing.

While the metadata standards and standards-compliant servers are helpful, consistency in local practices is still needed because inconsistent metadata adversely affect search results. If the metadata is incorrect, the resources in the IR are not adequately represented and will remain invisible to the users. Yasser (2011) extensive literature review identifies five categories of metadata problems: incorrect values, incorrect elements, missing information, information loss, and inconsistent value representation. Similarly, Alemneh (2008) and Shreves et al. (2005) explain that metadata problems occur when elements are not applied properly and when values are not consistently or accurately recorded. According to Dorman (2008), no open communication standard can compensate for metadata that lacks content or encoding standards. In the end, the visibility of digital items in an IR depends on the quality and richness of the descriptive metadata that content providers and digital curators provide.

**Conclusion**

The digital environment has introduced new resource types and new user expectations into the information landscape. In order to deliver a richer user experience and maximize visibility of digital resources, it is critical to have metadata that complies with standards, both in its completeness and its adherence to metadata creation standards. Consistency is the key factor in successfully managing different scholarly works in IRs and in making information more widely and easily available to users. Regardless of the repository software, metadata consistency plays a key role in describing and managing digital objects of different formats to effectively integrate the contents of IR into the existing services and collections. Although there are a number of contributing factors that affect digital resources visibility in IRs, it is the rich metadata that is consistently encoded that makes the digital items more discoverable.

**References**


A Participatory Framework for Evaluation Design

Diana S. Kusunoki
The iSchool at Drexel University
diana.s.kusunoki@drexel.edu

Aleksandra Sarcevic
The iSchool at Drexel University
aleksarc@drexel.edu

Abstract

Users and the information systems designed to support their needs and behaviors are becoming increasingly complex. Evaluators are tasked with designing evaluation methods that address the evaluation challenges of systems conceived through newer design principles, while also identifying issues and user perceptions in an efficient and effective manner. We argue that user involvement through structured formative evaluation design can fill the conceptual and procedural gaps between system design and system evaluation. This poster proposes a participatory framework for designing evaluation methods that can help designers and evaluators develop holistic approaches to design and evaluation by bringing greater detail, structure, and clarity to evaluation design. To determine the feasibility of this participatory evaluation design framework, we will integrate the system evaluation and evaluation design process into the system design process of information displays that support team situation awareness during trauma resuscitation.

Keywords: evaluation design, evaluation, participatory design, trauma resuscitation

Introduction

System design practice has progressed through several changes in perspectives on the role of the user, group work, and expanding contexts of use (Bannon, 1991; Bødker, 2006; Harrison et al., 2007). However, system evaluation is still grounded in traditional approaches (e.g., usability testing and heuristic evaluation) that were intended for implementation with users or experts individually in lab settings (Greenberg & Buxton, 2008). At the same time, systems are becoming increasingly dynamic and complex, requiring evaluation methods (i.e., questions, techniques, and instruments) that can adapt to the context in which they are used (Chilana et al., 2010). Evaluators may miss important issues if evaluation methods are not properly calibrated to the users, system characteristics, and design context. Moreover, without proper formative evaluation throughout system development, designers and evaluators risk wasting valuable resources and time with users.

Several challenges make it difficult to design contextually appropriate methods: (1) integrating system design, system evaluation, and evaluation design into one process; (2) coping with complex contexts; (3) reaching mutual understanding with users; (4) designing relevant evaluation methods; and (5) eliciting feedback and tacit knowledge from users (Kusunoki & Sarcevic, 2012). Current methods do not provide evaluators with guidelines for selecting or adjusting evaluation methods to ensure that both design and evaluation are grounded in the same theoretical foundations. A conceptual disconnect can also form if design and evaluation are treated as separate activities or even phases in the system development process. The goal of our research is to fill the conceptual and procedural gaps between design and evaluation by presenting an overarching framework that guides researchers in taking a holistic approach to system development through structured formative evaluation design. A participatory evaluation design framework can provide evaluators with more details about how to involve users in refining evaluation methods. It can also provide guidelines for how to implement targeted and streamlined evaluation and evaluation design activities during iterative system development.

Acknowledgements: We thank Drs. Andrea Forte, Michael Muller, Delia Neuman, and Susan Wiedenbeck for their valuable feedback.


Copyright is held by the authors.
A Participatory Evaluation Design Framework

To fill the conceptual gap, evaluators and designers need to merge their practices of understanding, designing, and evaluating. First, their practices of understanding user behaviors, perceptions, and contexts should provide a strong foundation for their practices of designing systems that support users in a particular context. Then, their practices of designing evaluation methods should align with their practices of designing systems. Finally, they must determine whether their understanding is properly reflected in their evaluation designs and the resulting system design.

Through the use of what we call a participatory evaluation design framework (Figure 1), evaluators can work with users to collaboratively determine what will be evaluated, why it should be evaluated, and how it will be evaluated. Participatory design can help evaluators develop methods that match user-generated evaluation design requirements. Evaluators can include users in the evaluation design process by collaboratively determining the usefulness, understandability, and appropriateness of the methods. Users help evaluators isolate and prioritize issues that must be addressed through the system design, which is especially useful in complex contexts. We emphasize user involvement without placing extra burden on users by carefully integrating system evaluation and evaluation design activities into the system design process to fill the conceptual and procedural gaps as described in the following sections.

Gather Requirements

While gathering system design requirements through techniques such as observations, participatory workshops, interviews, and surveys, evaluators can identify what users value and feel should be evaluated about the system for it to support their needs (Figure 1, gather requirements). The goal would also be to elicit the user perspective on topics such as values, issues, metrics for success, and feasibility of project. Time spent upfront on understanding users’ context, behaviors, perceptions, and specialized terminology will reduce time spent adjusting methods later.

Prototype

When the system prototype is being designed through rapid prototyping and participatory workshops based on the requirements gathered, evaluation methods could also be designed (Figure 1,
The goal would be to: 1) select and prioritize issues and values; 2) prototype methods that would elicit feedback; 3) decide which metrics will be used to determine the system’s success; and 4) identify potential uses of the evaluation results (Greene, 1987). Designers and evaluators would be responsible for maintaining the conceptual and technical integrity of the methods (Ayers, 1987; Garaway, 1995; Greenberg & Buxton, 2008; Greene, 1987).

**Test**

The methods are tested through user studies when conducting formative evaluation of the system design (Figure 1, test). Based on their experiences during prototype testing, evaluators determine from their perspective whether the evaluation prototype is functional, understandable, relevant, comprehensive, useful, and feasible to implement (Fowler, 1995; Groves et al., 2009). Evaluators then supplement their understanding with feedback from users through validation.

**Validate**

Users are asked to provide feedback about the functionality, understandability, relevance, comprehensiveness, usefulness, and feasibility of the evaluation prototype (Figure 1, validate). This can be accomplished through techniques such as debriefing, surveys, annotating instruments, and member checks (Fowler, 1995; Groves et al., 2009). Validation should follow testing immediately to ensure that the users’ perspectives are not affected by time delay. Member checks can be performed at a later time when initial findings are developed from evaluation testing and validation (Greene, 1987; Guba, 1981).

Based on user feedback about the evaluation method, evaluators may decide to iterate the prototype, test, and validate steps of the evaluation design process. Once all major user issues and concerns have been addressed through iterative evaluation, evaluators can elicit overall user perceptions about the system design and the evaluation design process. After concluding system development, rigorous summative evaluation will determine the technological performance of the system. Evaluators should periodically check that users are still satisfied because needs and technologies change rapidly over time.

**Specific Aims**

We have three specific aims that we plan to accomplish through implementing this participatory evaluation design framework:

- **Aim 1**: Fill the conceptual and procedural gaps between system design and evaluation.
- **Aim 2**: Design a framework for evaluators to systematically develop evaluation methods.
- **Aim 3**: Develop better questions, techniques, and instruments for eliciting users’ concerns, perceptions, understanding, and feedback on evaluation design.

In the following section, we discuss how we will implement this framework to structure the system development process of information displays that facilitate teamwork in the trauma resuscitation domain.

**Future Application in the Trauma Resuscitation Domain**

To determine the feasibility of this participatory evaluation design framework, we will integrate the system evaluation and evaluation design process into the system design process of information displays that support team situation awareness during trauma resuscitation. Trauma resuscitation is a highly complex, dynamic, and safety-critical medical domain. Healthcare providers from different disciplines, such as anesthesiology, nursing, and emergency medicine, form impromptu teams to perform life saving treatments on patients in critical conditions following the Advanced Trauma Life Support (ATLS) protocol (American College of Surgeons, 2008). Each trauma resuscitation is unique—different combinations of factors dynamically interact and contribute to the mechanism of the patient’s injury, symptoms, and reactions to treatments. Teams must rapidly collect and sift through extensive amounts of information from various sources to examine, diagnose, and treat patients (Sarcevic, 2008). Despite the urgent, high-risk, and complicated nature of the trauma resuscitation process, there are no information technologies present in the trauma room that aggregate and display contextual information to support team situation awareness and decision-making.
Additional Challenges in Trauma Resuscitation

A number of additional issues may arise during system design and evaluation due to nature of trauma resuscitation (Kusunoki & Sarcevic, 2012). Team members have various training backgrounds and levels of experience, and consequently may have conflicting ideas about what the design should support. Each team role also requires different pieces of information in different formats at different times. Teams are formed spontaneously and may not know each other or know each other’s work styles. Participant access is limited because healthcare practitioners are busy and working long, odd hours. Planning observations or user studies for live trauma resuscitations is unfeasible because they are infrequent and unpredictable. Privacy and security restrictions also make it difficult to receive clearance to view or record live trauma resuscitations. Moreover, testing unfinished systems in this environment is unsafe. Considering these challenges, system development needs to be flexible, efficient, and rigorous.

Meeting the Challenges

Techniques within the participatory evaluation design framework can be adapted to meet the challenges of the trauma resuscitation domain. Conducting simulations to test the system alleviates some issues with infrequency, privacy, and safety. Simulations also allow researchers to observe and videotape for detailed review. Recruiting participants from their appropriate disciplines to fill each role and randomizing user studies can reproduce the dynamic formation of diverse teams. Working with users early on will help ensure that the system design and evaluation methods are appropriate, ultimately saving everyone time by preventing costly repairs.

Conclusion

Regardless of the methods developed, overarching cohesion should exist among the principles and methodologies that support system design and evaluation. A participatory evaluation design framework can help designers and evaluators develop holistic approaches to system development. Designers and evaluators may potentially benefit from a structured framework for streamlining evaluation and evaluation design activities during system design. Future application in the trauma resuscitation domain will allow us to determine the feasibility and adaptability of this framework.

References


Practical Approaches and Proposed Strategies for Measuring Selected Aspects of Community-Based Broadband Deployment and Use

Lisandra R. Carmichael
Florida State University
lc09h@my.fsu.edu

Charles R. McClure
Florida State University
cmclure@lis.fsu.edu

Lauren M. Mandel
University of Rhode Island
lauren_mandel@mail.uri.edu

Marcia Mardis
Florida State University
mmardis@fsu.edu

Abstract

A series of studies of rural Florida community anchor institutions have concluded that existing national measurement practices for broadband penetration, adoption, and impact are often poorly defined, confusing, or inadequate to inform decisions about community broadband deployment and adoption. As a result, local broadband initiatives may be hindered by “measurement confusion.” We propose the Broadband Readiness Index (BRI) with a number of broadband readiness criteria to address this confusion and position local officials to better coordinate, deploy, and use broadband locally; demonstrate how improved high-speed broadband affects their communities over time; and sustain planning for continuous improvements of community broadband use.

Keywords: broadband penetration, broadband adoption, broadband measurement confusion, community broadband planning, broadband readiness index

Introduction

A series of recent studies conducted by the Information Use Management and Policy Institute (the Institute) at the Florida State University¹ revealed that local officials often found existing measures of broadband penetration, adoption, and impact confusing and impractical. This has hindered community broadband initiatives and resulted in investments with undocumented results. Realistic measurement options and solutions can alleviate local officials’ confusion about how to assess the broadband experience in their communities and at their community anchor institutions (CAIs), such as hospitals, libraries, and schools.

Selected Literature Disusing National Measures and Reporting

As a selected literature review revealed, many factors contribute to broadband measurement confusion, such as different speed tests using different methodologies (Bauer, Clark, & Lehr, 2010; FCC, 2010; Horrigan & Satterwhite, 2010; OECD, 2010), fixed versus wireless broadband, internet service providers (ISPs) offering different levels of services in the same communities, varying levels of data granularity (Kolko, 2010), and even users’ computer capabilities (Beard, Ford, Spiwak, & Stern, 2010;...
Benton & Williams, 2009; Connected Nation, 2011; Flamm, Friedlander, Horrigan, & Lehr, 2007; Smith, 2010; Whitt & Lampert, 2009). As a result, many community leaders have a difficult time clearly articulating their current or future broadband needs (Mandel, Alemane, & McClure, 2012). Assessing broadband status is complicated when CAIs and local officials are unaware of all the issues confounding broadband measurement. If decision makers are unaware of available broadband options and misunderstand their communities’ broadband needs, the communities they represent may have subpar broadband that will not meet their needs or assist their planning for the future.

Measuring Broadband Locally: An Example from Rural Florida Communities

This presentation builds on the overall conclusion from the North Florida Broadband Alliance (NFBA) and the Florida Rural Broadband Alliance, LLC (FRBA) projects that the lack of standardized measurements and indicators for a number of key broadband activities results in contradictory, complicated, and erroneous information, especially for practical use by local decision makers (McClure, Mandel, Saunders, Alemanne, Spears, & Bishop, 2011a and 2011b). Without accurate community-based deployment and measurement information, community leaders will find it quite difficult to justify costs and applications or show the impact that high-speed broadband has on a community’s health, education, economy, and overall quality of life. The purpose of the Broadband Readiness Index (BRI) is to help local community leaders understand how to measure broadband deployment and use to make informed decisions with regard to its adoption, deployment, and use.

The Institute conducted two concurrent needs assessments of broadband connectivity for the NFBA and FRBA projects in rural Florida. These projects shared a primary goal of bringing middle mile broadband infrastructure to Florida’s three Rural Areas of Critical Economic Concern (RACECs). Each RACEC is comprised of 6 to 14 rural, economically depressed counties. Each project included needs assessment, benchmarking, and onsite diagnostics, and employed a multi-method approach that included a web-based survey, focus groups, and onsite diagnostics at selected CAIs (Mandel, et al., 2012).

The select literature review suggests that ambiguities and confusion with national measures of broadband deployment and use also causes ambiguities and confusion at the community level. Indeed, the study team found that local community leaders’ eyes “glazed over” when detailing aspects of the National Broadband Plan about deployment, measurement, and related issues. The community leaders’ concern was solving broadband problems in their communities, not measuring national broadband deployment, speed, or jobs creation. From the FRBA and NFBA findings, the Institute developed an approach for measuring broadband readiness, deployment, adoption, use, and outcomes at the community level. The first step is for communities to assess CAIs readiness to implement broadband technologies. This approach utilizes a BRI that measures readiness according to nine criteria. The intent is to provide a practical, straightforward set of strategies to assist local CAIs, other organizations, and community opinion leaders to better access, deploy, and use broadband throughout their communities. After the Institute field-tests and refines these approaches, committed groups of community leaders can employ these strategies to improve broadband deployment and use in their communities.

Broadband Readiness Index

The BRI is based on nine criteria that take into account different situational factors under which CAIs operate and are meant to assess the readiness of an institution to adopt and utilize broadband. The degree to which an institution meets each criterion is assessed according to a three-point scale: high, somewhat, or low ability. The index requires both publicly available data and data that must be obtained from individual CAIs. Each criterion is broken down further into unique qualifiers intended to better predict the likely readiness of broadband adoption in the CAI:

Ability to change ISP. The ability to change ISPs is a critical situational factor affecting whether a CAI can adopt a broadband connection through a newly built middle mile or last mile network, fiber-to-the-premises, broadband over power line, or any other higher-speed technology than that to which the CAI already subscribes.

2 The RACECs are depicted at http://www.eflorida.com/FloridasFuture.aspx?id=2108
Available and trained IT staff. Available and competent IT staff in a CAI enable the other members of the staff to focus on their own tasks, provide administration with a reliable reference for CAI capabilities and needs, and conduct research on new applications that may provide a higher quality of service to clientele.

Existence and quality of internal network. A CAI’s internal network configuration is a major determining factor in its capability to adopt and fully utilize a broadband connection. Firewalls or poor configuration of workstation settings affect the ultimate speed of the connection reaching users and staff members as they complete tasks online.

Age of network and desktop equipment. The age of network and desktop equipment impacts a CAI’s ability to fully utilize a broadband connection. Ideally, a CAI replaces equipment on a 3- to 5-year schedule, but often CAIs purchase new equipment when critical components fail.

Sufficient funding. Funding for IT can be a fundamental problem for many CAIs. Many note that they are barely able to cover current costs and are unable to pay additional fees to break a contract or pay more for higher connection speeds or service quality. The structure of a CAI’s funding also influences how budgets are allocated. If the CAI receives technical support from a parent organization or funding primarily through grants, for example, it might not be able to allocate funds away from one area toward technology to buy new equipment in order to use new or upgraded broadband.

Administrative leadership. To successfully adopt broadband, CAI administrators must justify costs to boards and other governing bodies as well as to clientele. Sometimes administrators also must convince staff that using new technologies and changing current behavior are beneficial actions for the CAI.

Existence and quality of IT plan. To utilize broadband connections fully requires detailed short- and long-range technology planning. A high-quality IT plan enables a CAI to track past technology policy and equipment decisions and provides a reliable reference for administrators when confronted with contradictory information about services or equipment.

Administrative and staff interest in new technology applications. As previously stated, an administrator’s leadership and ability to persuade governing boards, parent organizations, and clientele are critical factors when considering a CAI’s readiness to adopt and utilize broadband. Fundamental for administrative leadership is an interest in finding new technology applications that facilitate new and improved service delivery methods. If the administration is thoroughly satisfied with the status quo, it is unlikely that it would be willing to put forth any significant effort to adopt broadband.

Demand from service population. Demand for innovative, high-quality services from a CAI’s service population is a critical component for justifying any change in service delivery methods or allocation of funding toward new technology. If the service population is uninterested or even hostile to new service applications available through broadband, it is unlikely that a CAI will attempt to change its current service offerings or delivery methods.

Conclusion

The BRI is proposed as a strategic approach that will improve successful community based broadband measurement, adoption, and deployment. The BRI is intended to be practical, easy to use, and comparable across different communities, and understandable to local decision makers. While the BRI has yet to be field-tested, it suggests broadband indicators and measures that community leaders can use to assess broadband readiness and monitor community broadband deployment and use. Indicators of broadband adoption and use in local communities include accessibility to the Internet, socio-economic development, and educational achievements. Finally, indicators of broadband applications include the availability of numerous service options, trends in use of e-government services, telemedicine, distance learning, and other broadband-enabled applications, and impacts on the community. There is a wide variety of possible metrics to measuring the impacts and outcomes of broadband, so “…a single model is unlikely to reveal all that should be explained about the impacts of broadband” (Holt and Jamison, 2009, p. 580). While the indicators proposed here may not be all-inclusive or measurable on a statewide or national level, they are applicable, reliable, and measureable indicators of broadband adoption in communities.
References


The Lifecycle Management of ETDs Project: Multi-Stakeholders National Partnership

Shannon Stark  
University of North Texas  
Shannon.stark@unt.edu

Daniel G. Alemneh  
University of North Texas  
Daniel.alemneh@unt.edu

Stephen Eisenhauer  
University of North Texas  
Stephen.eisenhauer@unt.edu

Abstract

The transition from traditional paper and microfilm formats to electronic theses and dissertations presents a number of significant challenges for academic libraries. To address these challenges, the UNT Libraries, together with their partners, are working on a collaborative project sponsored by an Institute of Museum and Library Services National Leadership grant. This poster describes the project objectives, possible outcomes, and proposed deliverables including a toolkit of guidelines, educational materials, and software tools.

Keywords: digital curation, ETD, lifecycle management, digital preservation

Introduction

In 2011, the Institute of Museum and Library Services awarded about a quarter million dollars to nine institutions and libraries to work collaboratively towards lifecycle management of Electronic Theses and Dissertations (ETDs). With virtually all theses and dissertations now being created as digital products, lifecycle management of these valuable intellectual documents is an important new responsibility of academic libraries. But it is a responsibility that unfortunately not all libraries have the infrastructure or resources to support.

The long-term management of ETDs requires not just the curation of electronic documents, but demands that consideration be given to policy creation, copyright, embargoes, formatting, metadata, and more. The big question is: How will institutions address the entire lifecycle of ETDs, ensuring that the papers acquired from students today will be available to future researchers?

This two-year project is the answer these nine institutions have devised. They will promote best curatorial practices to increase the capacity of academic libraries to reliably preserve ETDs by using a three-pronged approach that includes writing guidance documents, developing software tools, and creating educational materials—all of which will be available under an open access license at the conclusion of the project.

Participating ETD Curators

The following partners joined together to accomplish the stated objectives.

- University of North Texas (lead institution):
  - Dr. Martin Halbert, Principal Investigator
  - Shannon Stark, Assistant Coordinator
  - Dr. Daniel Gelaw Alemneh, Project Steering Committee Chair

Acknowledgements: Research contributing to this poster was funded by a Grant from the Institute of Museum and Library Services, awarded to the University of North Texas Libraries the Networked Digital Library of Theses and Dissertations, the Educopia Institute/MetaArchive Cooperative, and the libraries of Virginia Tech, Rice University, Boston College, Indiana State University, Pennsylvania State University, and University of Arizona. We are grateful to Martin Halbert, Dean, UNT Libraries and Katherine Skinner, Executive Director, Educopia Institute for constantly striving to improve the Library and Information Science profession.


Copyright is held by the authors.
Timeline for Deliverables and Workshop

The project began with the development phase in December 2011, with all members of the steering committee working virtually from their respective locations. During this first year, the focus has been on creating the eight guidance documents and micro-services, which will be completed by the end of 2012. In the second year, feedback on the deliverables will be gathered, and the workshop will take place. A timeline of these events can be seen in the following figure (Figure 1).

![Timeline of the project](image)

**Figure 1.** Timeline of the project

Project's Outcomes and Deliverables

A needs assessment process undertaken by the planning group for this project identified a cluster of informational resources required to advance the acquisition, organization, use, management, presentation, and preservation of ETDs. To address these needs, guidance documents are currently being written by experts in each specific area, based on their research and experiences.

Descriptions of Guidance Documents

Briefing on Access Levels and Embargoes of ETDs

One of the most contested topics in ETD program planning is the question of ETD embargoes and levels of access restriction. This document will directly address this difficult issue, which there is
Currently no consensus on regarding "good practices" in the field. This briefing will concisely lay out the issues and the different sides of this argument for those considering an ETD program, clearly articulating the pros and cons claimed by proponents for and against embargoes.

**Briefing on ETD Copyright Issues and Fair Use**

There is so much uncertainty and fear over the issue of copyright and fair use that it has in some cases prevented campuses from implementing ETD programs at all. This briefing will address publisher concerns and issues squarely. Again, the point of this briefing is to provide a concise summary of the relevant information for stakeholders, with documented citations.

**Guidelines for Implementing ETD Programs - Roles & Responsibilities**

This document will provide guidance for understanding roles and responsibilities, together with frank acknowledgements of the concerns that different stakeholders may have, and how ETD programs have addressed these issues as well as the ramifications of not specifying roles and responsibilities.

**Guidelines for Collecting Usage Metrics & Demonstrations of Value for ETD Programs**

This document will provide guidance for institutions concerning the critical issue of assessment of ETD usage, and how communicating such assessment metrics can demonstrate the benefits of the program to stakeholders. Practical examples of how to document and convey usage metrics will be provided.

**Overview of Formats, Complex Content Objects, and Format Migration Scenarios for ETDs**

Additional issues of great concern to many current and prospective ETD programs are how to determine what formats they should accept, and how to manage on an ongoing basis the increasingly complex ETDs that are now being produced by students. This document will seek to set forth the relevant information and issues, and to provide scenarios for managing and migrating complex ETDs over their extended lifecycle.

**Overview of PREMIS Metadata & Lifecycle Event Record-Keeping for ETDs**

Another issue revealed in the needs assessment process was that most institutions do not have workflows and systems in place to capture the appropriate levels of metadata needed to manage ETDs over their lifecycle, often because of a lack of awareness of the ramifications of not maintaining such information. An overview of these critical issues will be provided, not just at the point of ingestion, but subsequently, as ETDs often have transitional events in their lifecycle (embargo releases, redactions, etc.).

**Guide to ETD Program Cost Estimation and Planning**

One of the most frequently asked questions identified in the CNI and MetaArchive surveys pertains to the costs and process for setting up an ETD program. This guide will provide succinct information on costs and planning, laying out the critical paths that many ETD programs have charted, together with advice about strategies.

**Guide to Options for ETD Programs**

Many institutions are delayed in ETD program planning simply because they do not have a clear understanding of the range of options to consider in implementing an ETD program. Like the other documents in this series, this guide will be a descriptive (rather than prescriptive) information resource.
that seeks to explain the relevant decisions institutions must make and to clarify the pros and cons of different options. This document will reference and integrate all the other documents in the series.

Descriptions of Tools

The use of software tools to analyze ETD content in various ways is an essential part of the repository workflow. These tools help analyze the formats of digital files, record information about their formatting and validation, and scan them for computer viruses. One of the goals of this project is to develop a common protocol for allowing any institutional repository to use these tools effectively and with minimal effort.

Despite the usefulness of these tools, integrating them into an existing institutional workflow or ETD management system can be challenging and require large amounts of code written specifically to interact with each individual tool. We hope to simplify this challenge by creating a specification for scripts that can bridge the gap by handling the direct interactions with these tools, and making their output usable in a clean, uniform manner (Figure 2). This will greatly reduce the development burden of institutions wishing to integrate these tools.

![Figure 2. Tools development for ETDs lifecycle management project.](image)

Descriptions of Educational Materials and Associated Workshop

During the second year of the project, a set of educational materials will be created on the theory and practice of lifecycle management of ETDs—which will be made freely available—and give a workshop utilizing these materials. These materials will reference the guidance documents, while the workshop will offer hands-on lifecycle tools training. The educational materials will include syllabi, training handouts and exercises, and PowerPoint presentations.

Summary

Theses and dissertations are predominantly produced in digital format today, and it is the academic libraries’ responsibility to meet the challenge of managing these documents throughout their lifecycle. This project provides three key resources to help practitioners succeed. The guidance documents will provide libraries with a much needed core set of concise advisory documents to address practical implementation considerations in ETD programs. Although necessarily limited in scope, the software tools address key functions that virtually all ETD programs must provide, and will be available
via a free and open source software license for any academic library to utilize. And lastly, the educational materials and workshop will enable this project to directly reach practitioners seeking information and tools relevant to establishing or updating ETD programs at their institutions.

References

Implementing an Open Source EMR in a Nursing Informatics Course

Nina Multak  
nlm38@drexel.edu

Emad Khazraee  
sk827@drexel.edu

Michelle Rogers  
mlr92@drexel.edu

Prudence Dalrymple  
pwd28@drexel.edu

Drexel University

Abstract

Electronic medical record (EMR) systems are being implemented widely in hospital and office based settings. This technology is becoming a standard documentation modality for inpatient and outpatient clinical settings. While practicing clinicians have learned to use these systems with the benefit of institutional or practice based training, the educational benefit of providing EMR training to students prior to their exposure to the clinical environment has not been widely explored. This preliminary research evaluated multiple EMR systems for educational use in a nursing school curriculum. EMR training including classroom lectures, computer based training sessions and practice with interdisciplinary medical simulation are incorporated in this educational program, providing students access to an EMR system before using one with actual patients. The ADDIE model of systemic instructional design was utilized for this study. Additionally, deliberate practice, an evidence-based construct grounded in information processing and behavioral theories of skill acquisition and maintenance guided the design of this study. Preliminary results reveal a high level of learner interest and satisfaction with the training methodology. Further exploration of EMR systems and the impact this component has on nursing education and practice are planned.

Keywords: education, EMR, simulation, documentation

Introduction

The Institute of Medicine Committee on Data Standards for Patient Safety reports evidence that improved clinical efficiencies, patient safety, and quality care outcomes can be promoted by having an effectively structured and implemented EMR1. Medical practices and hospitals are discovering the importance of adequate training and best methods for bringing everyone up to speed to ensure a successful transition to EMR’s1. Electronic health record systems have four core functions which include: health information and data, results management, order entry and support, and clinical decision support. Other functions of the electronic health record system include patient support, electronic communication, administrative support, reporting and population management2. Information provided by health information technology (HIT) to clinicians, will assist them with patient support decisions, where accuracy and speed are vital to patient outcomes. The EMR can identify the quality and quantity of care rendered, and serves as a communication tool to aide multidisciplinary healthcare team members. Exposure to electronic health technologies will help providers increase their comfort level, accuracy and speed using these systems. Electronic health technologies such as the electronic medical record can be used to train student clinicians effectively in the management of patients. While many institutions and physician practices have implemented EMR systems, there is only a scant amount of literature about training clinicians in electronic documentation.

Expectations for EMR’s suggest that it will improve patient safety. In institutions with well-developed, longstanding HIT, evidence suggests that improvements have been made3. Electronic clinical documentation systems providing electronic capture of clinical notes and data exchange. The majority of certified EMR's allow for customized templates, permitting practices to gather relevant information for their work on a single screen, while populating the information into the patients' medical record. EMR’s
require careful organization for ease of data input and content retrieval. This preliminary study evaluates an educational training program which provides student clinicians an opportunity to record, store, retrieve and modify information in an EMR in a simulated clinical environment.

**Materials and Methods**

Healthcare students in an east coast university in the United States were initially introduced to documentation skills in a lecture format. An emphasis of the lecture was the importance of good documentation as the primary means of chronologically recording patient care to facilitate communication among professionals involved in a patient’s care. Legal considerations of medical documentation were also provided. Practice opportunities were provided through an Excel based EMR which was designed and implemented into the classroom and implemented into established medical simulation cases using standardized patients and high fidelity mannequin simulators. The case scenarios involved healthcare provider students from various disciplines. The EMR was substituted for a traditional paper based chart.

Individualized interviews were completed with 12 healthcare students - six students interviewed were receiving clinical training in an inpatient setting and six students interviewed were receiving training in outpatient settings. Students supported or strongly supported the addition of hands-on training in the curriculum. Feedback from students suggested that utilization of a more realistic EMR system would enhance their ability to use this training in an effective manner. To achieve this end, we have done a survey among EMR solutions. Enterprise EMR solutions are very expensive to implement. Given the exploratory nature of this project, we focused on open source solutions for three main reasons: Lower initial cost of implementation, flexibility to modify the solution to best serve educational goals of the project, and opportunity to receive support from a wider development community as opposed to just one vendor.

Since there is little information available on the evaluation of the existing open source EMR solutions for educational settings, we collected information about existing open source EMRs and evaluated them to find the best candidates. Table 1, summarizes the result of this survey completed in February 2012. Finally, we consider the important factors such as acceptance of the solution in the healthcare community, existence of documentation, inpatient and outpatient support, frequency of updates, and community support. Finally, Open Vista was selected as the platform for the educational EMR. Open Vista is developed by Medsphere based on the widely used VistA EMR developed by the Department of Veteran Affairs over past two decades. There are a family of VistA based solutions but among them Open Vista provides a redesigned user interface and the possibility to use both community support and commercial support. VistA has been used for a long time in many venues and it is known as a reliable solutions. In addition, it supports both inpatient and outpatient features as well as wide variety of procedures required for the educational purposes.

**Preliminary Results**

During four terms of teaching a nursing informatics course, it was found that students asked for hands-on experience with a real world enterprise EMR. Even the use of an Excel-based EMR did not satisfy their desire to experience an EMR they may be exposed to in their future practice. The first experience of students interacting with OpenVista both via client side application and hand-held devices was very promising. Students were able to experience a rule-based alarm system and documentation using standard terminologies.

Faculty observation during the simulation sessions revealed a 1.25-4.5 minute interval from the time of clinician introduction to the patient to the onset of documentation (n=34). Duration of documentation varied due to details of simulation cases. Interviews with student health care providers revealed a strong interest (100%) in receiving additional training in EMR documentation and 88% supported the application of EMR documentation with medical simulation scenarios.
<table>
<thead>
<tr>
<th>Software</th>
<th>Type</th>
<th>Programming Language</th>
<th>Database</th>
<th>Platform</th>
<th>Access</th>
<th>Website</th>
<th>License</th>
<th>Latest Version</th>
<th>Last Release Date</th>
<th>Documenta tion</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenVistA</td>
<td>Inpatient/Outpatient</td>
<td>MUMPS/GT.M</td>
<td>MUMPS/GT.M</td>
<td>Unix/Linux</td>
<td>Server/Client</td>
<td>Link</td>
<td>AGPL/LGPL</td>
<td>1.5 SP6</td>
<td>05/24/11</td>
<td>Yes+Online Demo+Appliance</td>
</tr>
<tr>
<td>World VistA</td>
<td>Inpatient/Outpatient</td>
<td>MUMPS</td>
<td>GT.M/Cache</td>
<td>Cross-platform</td>
<td>Server/Client</td>
<td>Link</td>
<td>GPL v2</td>
<td>1.0</td>
<td>03/26/11</td>
<td>Yes+Online Demo+Appliance</td>
</tr>
<tr>
<td>Astronaut</td>
<td>Inpatient/Outpatient</td>
<td>MUMPS</td>
<td>MUMPS/GT.M</td>
<td>Unix/Linux</td>
<td>Server/Client</td>
<td>Link</td>
<td>GPL v3</td>
<td>0.9 Beta</td>
<td>09/30/09</td>
<td>Some</td>
</tr>
<tr>
<td>ClearHealth</td>
<td>Inpatient/Outpatient</td>
<td>PHP</td>
<td>MySQL</td>
<td>Cross-platform</td>
<td>Server/Client</td>
<td>Link</td>
<td>GPL</td>
<td>3 Beta</td>
<td>Aug 2011</td>
<td>Yes</td>
</tr>
<tr>
<td>WebVista</td>
<td>Inpatient</td>
<td>PHP</td>
<td>NA</td>
<td>Cross-platform</td>
<td>Web-based</td>
<td>Link</td>
<td>LGPL/GPL L2/AGPL</td>
<td>Beta</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>vxVistA</td>
<td>Inpatient/Outpatient</td>
<td>MUMPS</td>
<td>MUMPS/GT.M</td>
<td>Cross-platform</td>
<td>Server/Client</td>
<td>Link</td>
<td>EPL</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>OpenMRS</td>
<td>Ambulatory</td>
<td>Java</td>
<td>MySQL</td>
<td>Cross-platform</td>
<td>Web-based</td>
<td>Link</td>
<td>-</td>
<td>1.8.3 stable 1.9 Beta</td>
<td>12/13/11/12/22/11</td>
<td>Yes+Online Demo+Appliance</td>
</tr>
<tr>
<td>OpenEMR</td>
<td>Ambulatory</td>
<td>PHP</td>
<td>MySQL</td>
<td>Cross-platform</td>
<td>Web-based</td>
<td>Link</td>
<td>GPL</td>
<td>4.1.0</td>
<td>09/23/11</td>
<td>Yes+Online Demo+Appliance</td>
</tr>
<tr>
<td>OSCAR</td>
<td>Inpatient/Outpatient</td>
<td>Java/JSP</td>
<td>MySQL</td>
<td>Cross-platform</td>
<td>Web-based</td>
<td>Link</td>
<td>GPL</td>
<td>10.12-1.2 final 11-1.1alpha</td>
<td>09/17/11/01/1512</td>
<td>Yes+Online Demo</td>
</tr>
<tr>
<td>Patient OS</td>
<td>Ambulatory/Inpatient</td>
<td>Java</td>
<td>PostgreSQL/QL</td>
<td>Cross-platform</td>
<td>Server/Client</td>
<td>Link</td>
<td>GPL</td>
<td>1.3</td>
<td>08/03/11</td>
<td>Some</td>
</tr>
<tr>
<td>GNU Health</td>
<td>Ambulatory/Inpatient</td>
<td>Python</td>
<td>PostgreSQL/QL</td>
<td>Cross-platform</td>
<td>-</td>
<td>Link</td>
<td>GPL</td>
<td>1.4.3</td>
<td>01/30/12</td>
<td>Some</td>
</tr>
<tr>
<td>GNUmed</td>
<td>Ambulatory</td>
<td>Python</td>
<td>PostgreSQL/QL</td>
<td>Cross-platform</td>
<td>Server/Client</td>
<td>Link</td>
<td>GPL</td>
<td>Client 1.1.0 Server 16.0</td>
<td>10/27/11</td>
<td>Yes</td>
</tr>
<tr>
<td>THIRRA</td>
<td>Ambulatory</td>
<td>PHP</td>
<td>PostgreSQL/QL</td>
<td>Cross-platform</td>
<td>Web-based</td>
<td>Link</td>
<td>MPL</td>
<td>0.9.12</td>
<td>11/23/11</td>
<td>Yes+Online Demo</td>
</tr>
<tr>
<td>FreeMED</td>
<td>Ambulatory</td>
<td>PHP</td>
<td>MySQL</td>
<td>Cross-platform</td>
<td>Web-based</td>
<td>Link</td>
<td>GPL</td>
<td>0.8.5</td>
<td>11/04/11</td>
<td>Some+Online Demo</td>
</tr>
<tr>
<td>Care2x</td>
<td>Inpatient</td>
<td>Java/PHP</td>
<td>MySQL/PostgreSQL/QL</td>
<td>Cross-platform</td>
<td>Web-based</td>
<td>Link</td>
<td>GPL/LGPL</td>
<td>2.6.27</td>
<td>01/03/2012</td>
<td>Yes+Online Demo</td>
</tr>
</tbody>
</table>
Conclusion

Training nursing and allied healthcare students to document in EMR’s before practicing clinically and assessing these individuals for competency will likely contribute to reduction of documentation based errors. Exploration of the implementation of EMR training using the Open Vista system is currently being explored with current Nursing Informatics courses at this institution. This educational program which incorporates reflection in action to teach critical thinking and clinical judgment can be used as a new learning analytic for the measurement, collection and reporting of information about these learners and their progress

References


Haugen H. The advantages of simulation training. Health Mgmt Tech (2012) Apr; 18
Toward a Mesoscopic Analysis of the Temporal Evolution of Scientific Collaboration Networks

Syed Ishtiaque Ahmed
sa738@cornell.edu
Department of Information Science, Cornell University

Scott Allen Cambo
sac355@cornell.edu

Carl Lagoze
clagoze@umich.edu

Theresa Velden
tvelden@umich.edu
School of Information, University of Michigan

Abstract

This poster reports on our latest results in a multiyear project that employs a mixed network analytic and ethnographic approach to understand the factors underlying field-specific attitudes towards openness and sharing of scholarly data. We report initial results of adding a temporal dimension to an analysis of scientific collaboration networks that provide evidence for comparative study of community structures and collaboration patterns across scientific fields. The addition of a temporal dimension to the analysis allows us to study the dynamic processes involved in the evolution of a scientific community and to determine field specific patterns. Further, it improves the accuracy with which the internal structures of scientific collectives can be resolved. This ongoing work advances an ethnographically grounded approach to the mesoscopic analysis of collaboration networks. Supported by ethnographic insights, we can connect mesoscopic network features to notions of research groups, group leadership and implied seniority, inter-group collaboration, between group migration, and ephemeral one-off exchanges. Eventually, a mesoscopic perspective should allow us to significantly improve the validity of models to explain network evolution.

Keywords: co-author networks, mixed methods, temporal evolution, scientific collaboration, community structures

Introduction

This poster reports on our latest results in a multiyear project that employs a mixed network analytic and ethnographic approach to understand the factors underlying field-specific attitudes towards openness and sharing of scholarly data. We report initial results of adding a temporal dimension to an analysis of scientific collaboration networks that provide evidence for comparative study of community structures and collaboration patterns across scientific fields. The addition of a temporal dimension to the analysis allows us to study the dynamic processes involved in the evolution of a scientific community and to determine field specific patterns. Further, it improves the accuracy with which the internal structures of scientific collectives can be resolved. The results can then be used to guide the strategic sampling of field sites for comparative ethnographic field studies.

This work aims at advancing an ethnographically grounded approach to the mesoscopic analysis of collaboration networks (Velden, Haque, & Lagoze, 2010; Velden & Lagoze, 2012). Supported by ethnographic insights, we can connect mesoscopic network features to notions of research groups, group leadership and implied seniority, inter-group collaboration, between group migration, and ephemeral one-off exchanges. Previous work has oftentimes conceptualized co-author nodes as autonomous actors

Acknowledgments: National Science Foundation Grant No OCI-1025679, International Fulbright Science and Technology Fellowship for S.I. Ahmed.


Copyright is held by the authors.
driven by individualistic mechanisms such as preferential attachment, ignoring the actual social composition of research collectives and the various socially distinct processes contributing to global network growth and densification. Eventually, a mesoscopic perspective should allow us to significantly improve the validity of models to explain network evolution.

**Methods**

We are developing an open source code base (http://github.com/tvelden/communities) that allows us to flexibly generate co-author networks following different time-slicing schemes: ‘accumulative’ for tracking the accumulative growth of the network, and ‘sliding’ for generating a dynamic view of the evolution of network structures by considering only publications in a specific time window. This sliding window can move across the entire time range covered by the available data. We have integrated into the code methods that support the mesoscopic analysis of networks, such as the network clustering code by (Rosvall & Bergstrom, 2008) and our own implementation of a node classification algorithm for clustered networks by (Guimera, Sales-Pardo, & Amaral, 2007). The latter classification scheme allows us to distinguish types of nodes by their structural embedding into their surrounding co-author cluster as well as by their out of cluster connectivity. For example, hub nodes extracted from our networks by this classification scheme can be identified as research group leaders (Velden et al., 2010).

**Data**

We have developed a lexical query to extract from the Web of Science (WoS) of Thomson Reuters the publication output of two fields in the physical and chemical sciences between 1991-2010, one in synthetic chemistry (field 1), and one at the boundary of physics and physical chemistry (field 2). An important step is the cleaning of data. To improve the accuracy of the co-author networks we apply an author disambiguation algorithm (Velden, Haque, & Lagoze, 2011). We further use a statistical approach to define hyper-authorship in a data set-specific way and use it to exclude a small set of papers (1-3%) that are not representative of the research style in the long-tail science fields that we study here. A manual analysis finds that in many cases those hyper-authorship papers represent out-of-scope papers that the lexical query mistakenly captured. In a few cases we also find large-scale collaborations that contribute to the specific field we study, but represent only a marginal sub-community within the field. Finally, we exclude authors who have co-authored only a single paper. About two-thirds of authors are removed in this step. We found that metrics for the global network topology were not affected by this latter reduction step. The reduced data is much more manageable for analysis and visualization purposes, such as the visualization of the giant component of the network of field 2 in figure 1.

![Figure 1](image-url)  
*Figure 1*. Network structure of giant component of field 2 at different stages of its (accumulative) evolution. Between 1992 (a) and 1993 (b) the initial core of the giant component is formed and the network diameter suddenly increases in size. In 1995 (c), even though the network has been further growing in size, the network diameter decreases, indicating the increasing densification of the network over time.
Initial Findings

We calculate global network metrics (table 1) and compare them to the characteristics of co-author networks investigated by (Bettencourt, Kaiser, & Kaur, 2009) to confirm that our network data represent scientific fields within a common range of topological characteristics and that we are not dealing with extreme outliers. We find that the global metrics conform broadly to the characteristics of other fields. For both our fields, the scaling parameter for network densification is similar e.g. to the field of carbon nanotubes and in accordance with a non-pathological fields of a community of researchers that share concepts and techniques (Bettencourt et al., 2009).

Table 1
Basic Network Properties and Global Metrics for Preprocessed Data

<table>
<thead>
<tr>
<th></th>
<th># papers</th>
<th># authors</th>
<th>edges (weighted)</th>
<th>diam. giant</th>
<th>size giant (nodes)</th>
<th>size giant (edges)</th>
<th>densification (scaling param.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field 1</td>
<td>12,641</td>
<td>13,397</td>
<td>58,375</td>
<td>ca. 35</td>
<td>62.3%</td>
<td>76.7%</td>
<td>1.14</td>
</tr>
<tr>
<td>Field 2</td>
<td>56,122</td>
<td>60,457</td>
<td>315,491</td>
<td>ca. 40</td>
<td>65.1%</td>
<td>84.2%</td>
<td>1.18</td>
</tr>
</tbody>
</table>

We then focus on the question of how new authors join the network and the role of preferential attachment, initially by replicating for our data the analyses done by (Abbasi, Hossain, & Leydesdorff, 2012) on a data set for the field of ‘steel structure’ (1999-2008) research and by (Milojevic´, 2010) on a data set on ‘nano science’ (2000-2004). Abbasi et al. suggest that betweenness centrality is the driver of preferential attachment in the evolution of research collaboration networks, more so than degree centrality, which traditionally has been the focus of preferential attachment models. They observe for their data set that the betweenness centrality of existing nodes in a co-author network correlates more strongly with the number of new authors they attract and link to than degree centrality. However, we consider the 0.23 to 0.32 Spearman correlation strengths that they report as relatively low, having limited explanatory value for the variation observed in the number of links existing authors form with new authors. Our data show an even lower correlation (between 0.1 and 0.2 for field 1, and 0.05 and 0.15 for field 2), and the correlation values we obtain for degree centrality and betweenness centrality are almost the same. We have conducted an additional analysis focused specifically on hub nodes (i.e. research group leaders) and the number of new authors that link to them. We find that the correlation between centrality of a hub node and number of new authors linking to them vanishes, suggesting that new authors entering the field do not discriminate their attachment to hub nodes by the respective centrality score of the hub node in the network. Hence we cannot corroborate the finding of a dominance of betweenness centrality over degree centrality as driver of the evolution of collaboration networks. Further, the low correlation values contradict claims that preferential attachment based on network centrality plays a major role in explaining attachment dynamics of new authors.

As pointed out by Milojevic power law scaling of the distribution of number of collaborators (associated with a hypothesized preferential attachment mechanism at work), is not the dominant feature characterizing such distributions in co-author networks. Instead, the majority of authors (88% in the 2000-2004 data set of nano-science publications studied by Milojevic) are included in the log-normal hook of the distribution. Milojevic interprets the hook and its peak as suggestive of a characteristic mode of collaboration corresponding to the typical number of collaborators needed in a research field to produce a publishable result. Our data for both fields display the same log-normal hook feature with a peak at 2 collaborators. These peak values are slightly smaller than those in nano-science subfields that could be comparable to our fields. This could be due to differences in preprocessing of the data. We have started analyzing the extent to which these features persist if we consider only specific (Guimera) classes of nodes.

Finally, we are investigating how the network evolves at the level of network components. One phenomenon in our data that caught our attention has been the temporal oscillation of the size of the second largest component. Whenever its size drops, this indicates a ‘feeding event’ in which its nodes join the giant. We have checked whether the second largest component acts as a major ‘staging ground’ for nodes to join the giant. Testing this hypothesis we find that after 20 years only 8% (field1) or 2% (field...
2) of the nodes in the giant component have at some point in the past been members of the second largest component. The largest annual influx of nodes to the giant component are new nodes entering the network (typically 50-70% for both fields). For field 1, however, we find a potentially interesting 3-year phase of successive mergers of the giant with the second largest component such that in 1999 eventually 28% of all nodes in the giant component have passed at some point through the second largest component.

Conclusions

We suggest that attempts at explaining the dynamics of network growth need to distinguish more carefully between the different types of nodes and social processes underlying co-authorship collaborations. We expect valuable insights into the evolution of collaboration networks in scientific communities and field specific collaboration patterns form an ethnographically grounded and time-sensitive analysis of collaboration networks. We here focus on co-author networks, however in future work we anticipate to include in the temporal analysis of layered citation and co-author networks for the mapping of community structures within scientific fields.

We also note that our experiences with the replication of other authors’ results revealed a number of critical issues that underline the potential benefit of an open data approach, allowing routine sharing of the data sets underlying published analyses, for developing a strong reliable empirical base for field comparisons.

References


Label Annotation through Biodiversity Enhanced Learning

P. Bryan Heidorn  
University of Arizona  
heidorn@email.arizona.edu

Qianjin Zhang  
University of Arizona  
zhqjn@gmail.com

Abstract

The LABELX (Label Annotation through Biodiversity Enhanced Learning) is an extension of the HERBIS NLP system reported previously (Heidorn & Wei, 2008). The objective of the system is to formaly structure output from Optical Character Recognition (OCR) of the highly variable labels of natural history museum specimens. OCR errors are common in the OCR output. Genus and species names are particularly prone to errors. Records are preprocessed using a fuzzy-match algorithm to find and replace genus and species names, including those with OCR errors, and replace those with a constant token. Integers and strings that begin with Alphabetic characters and end with numbers are also replaced with tokens. LABELX generates structured XML data and RDF and makes corrections to OCR errors in some fields. The main algorithm is a Hidden Markov Model (HMM). This poster reports an enhancement to the previous system with a larger data set.

Keywords: OCR, parsing, semantic markup, machine learning

Introduction

LABELX is a set of programs designed to process herbarium label data into a format that can be easily ingested into museum databases and distributed in standard formats over the Internet. “T(t)wo thirds of the collections have less than 75% of their collection data online. Images are less accessible than other data: more than 90% of collections have less than 10% of their holdings represented online in digital images.” (Skog et al., 2009). Collections of plants have existed since the beginning of civilization. In the 1540s (Pavord, 2005) the collection of plants took on what we would call scientific organization. According to the Index Herbariorum there are about 3,400 herbaria in the world (Thiers, continuous). The labels on specimens are a rich source of information about not only the name of plants but their historic distribution and environmental requirements. Early labels were of course handwritten. In the 20th century the labels begin to be typed but even then scientific names are handwritten. Even within typewritten labels there is a great variety of formats. Formats are created by individual collectors or by curators of particular collections or by the project managers for biodiversity surveys or other projects.

For type-written labels Optical Character Recognition (OCR) is a potentially valuable method to convert label images into machine readable UTF-8 format. In the resulting text the elements of the label will still be in varying orders. Labeling the elements of the label is the focus of the current study. Unfortunately, OCR on labels are prone to errors. Our objective in this study is to properly label the elements of these labels in spite of the OCR error rate. These OCR errors have a number of sources. The fonts being used in labels are often non-standard and vary within and across labels. Typewriter ribbons were not always of high quality so there are partial letters. Typewriters sometimes had loose pinions so letters do not line up with one another. Many of the words that appear in labels are not in standard OCR dictionaries so the algorithms built into many OCR engines are not able to successfully guess full words based on the context of remaining characters.

Acknowledgement: Funding provided by National Science Foundation, #0956271 to Bryan Heidorn. We thank Steven Chong for assistance in gathering data and RDF output.


Copyright is held by the authors.
Sample Labels

The labels in the current sample come from two sources. The first set of 130 labels was from the Yale Herbarium similar to those reported in 2008 (Heidorn & Wei, 2008). The original Yale OCR collection was relatively small with just a couple of hundred items and the two-thirds with the best OCR were selected for analysis. 200 new labels were added from a research collection from the University of Alaska, Museum of the North digitization project. The new labels from the Museum of the North are part of 55,000 labels we have on file. The selected records are the first from that file excluding records with poor OCR. In the current project we identified 55 different element types in the labels. The current analysis includes 40 elements because the remaining items appeared less than 20 times in the collection making machine learning methods ineffective.

The Algorithms

If the museum labels were highly regular it would be possible to write regular expressions to extract information for the labels. The labels are not that regular but there are some common patterns that can be exploited. Some label elements include human readable tags that were used to tell humans where to type content on a label. Some labels contain the string “S/N:” to signal scientific name. Others contain the string “Name:” and others contain no indication the scientific name follows except the general location on the label. Sometimes this is right after the collector number. A scientific name is composed of predictable parts so Genus is followed by species, which is usually followed by authority. To exploit these regularities two main algorithms were used in this study. The first is Hidden Markov Models (HMM). The performance of HMM in this domain has been reported elsewhere (Heidorn & Wei, 2008). This algorithm which is used in genetics and other domains identifies sets of orderings of elements that best matches previously seen orders of elements (Frasconi, Soda, & Vullo, 2003).

The contribution in this study is the use of fuzzy matching to identify genus and species using an authority file. Scientific names and collector names are difficult for HMM because they are relatively rare. A training set may have several hundred entries but each name may appear only once. We could preload the HMM with many thousands of names from an authority file but this leads to performance problems. Instead, we replace Genus and Species with fixed tokens. For scientific names we use the International Plant Name Index (IPNI). IPNI is a database of the names and associated basic bibliographical details of seed plants, ferns and lycophytes. IPNI is a dynamic resource, depending on direct contributions by all members of the botanical community (IPNI, 2012). Collector names, Genus names and species names are not always unique and may sometimes be other words in English or other languages such as “rose”. Simple substitution leads to false positives. OCR errors leads to false negatives. This fuzzy algorithm takes these conditions into account in most cases. If both the genus name and adjacent species name match the entries in IPNI they are replaced with a constant “BiSciColGenus” and “BiSciColSpecies” respectively. If only the genus name is found in IPNI the algorithm creates a list of just species of that genera from IPNI. It then measures the Levenshtein distance between the unknown word and the species names. If a threshold of .8 is met the string is assumed to be a misspelling of the species and is substituted with the string “BiSciColSpecies”. The match from IPNI is recorded as the correct spelling. In the following example the genus has an OCR error. The algorithm finds the species as an exact match and then recognizes the genus.

1) The original misspelled full genus names (H1126562): AGR CSTIS EXARATA Trin.
2) The substituted result: BiscicolGenus BiscicolSpecies Trin.
   The spelling correction: AGROSTIS EXARATA Trin.

Conversely, if an exact match is found for a species name but the string before that exact species match is not a genus, the algorithm collects all genus names that are valid for that species name from IPNI. The Levenshtein distance is calculated between the unknown prior string and the members of this genus list. If any meet the threshold the unknown string is replaced with the constant “BiSciColGenus.” The match from IPNI is recorded as the correct spelling. The following table (Table 1) lists the fields with performance changes of greater than +/- 3%. Species author names increase because of ordering effects detected by the HMM because of improvement of Genus and species performance. Labels of fields are excluded since they are not placed in databases. Deleterious effect of number substitution are seen in altitude and date fields. These will be countered by adding explicit matching for common date formats.
Habitat accuracy decreases because scientific names sometimes appear in habitat descriptions as associated species. The loss of common name accuracy is attributable to confusion with collector names.

Table 1  
Change in F-Score

<table>
<thead>
<tr>
<th>Head</th>
<th>Control</th>
<th>Treatment</th>
<th>%Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>.709</td>
<td>.747</td>
<td>+3.84</td>
</tr>
<tr>
<td>Genus</td>
<td>.771</td>
<td>.812</td>
<td>+4.07</td>
</tr>
<tr>
<td>Species</td>
<td>.768</td>
<td>.806</td>
<td>+3.81</td>
</tr>
<tr>
<td>Species author</td>
<td>.699</td>
<td>.758</td>
<td>+5.93</td>
</tr>
<tr>
<td>Collector name</td>
<td>.760</td>
<td>.796</td>
<td>+3.60</td>
</tr>
<tr>
<td>Collection number</td>
<td>.589</td>
<td>.670</td>
<td>+8.16</td>
</tr>
<tr>
<td>Barcode</td>
<td>.463</td>
<td>.919</td>
<td>+45.56</td>
</tr>
<tr>
<td>Latitude/Longitude</td>
<td>.232</td>
<td>.343</td>
<td>+11.07</td>
</tr>
<tr>
<td>Habitat</td>
<td>.699</td>
<td>.665</td>
<td>-3.41</td>
</tr>
<tr>
<td>Determiner name</td>
<td>.746</td>
<td>.683</td>
<td>-6.32</td>
</tr>
<tr>
<td>Common name</td>
<td>.762</td>
<td>.672</td>
<td>-8.99</td>
</tr>
<tr>
<td>Altitude</td>
<td>.577</td>
<td>.498</td>
<td>-7.84</td>
</tr>
<tr>
<td>Determination</td>
<td>.739</td>
<td>.656</td>
<td>-8.28</td>
</tr>
<tr>
<td>Date</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

Fuzzy Matching techniques such as Levenshtein distance cannot be applied directly to all strings in errorful OCR of museum labels because of performance limitations and the false positive matches. However, OCR corrections are possible by evaluating context of the words before or after a potential fuzzy match. This can result in improved semantic labeling of neighboring items when using HMM. Substitution of numbers and alphanumeric strings with high variability can also improve semantic classification of items on labels at the cost of classification of some dates. Minor modifications of the substitution algorithm for recognition of standard date patterns should help overcome this limitation.

References


Surveillance of Audience Labor Using New Media: Three Innovations of Television Broadcast Networks

Brian L. Griffin
University of Toronto
brian.griffin@mail.utoronto.ca

Abstract

In response to the proliferation of channels, the growth of the Internet, and the availability of time-shifting and advertisement-blocking technologies, how are television broadcast networks using new media innovations to efficiently exploit and sell audience labor to advertisers? Drawing upon insights from political economy approaches to audience labor, this poster outlines three new media strategies television broadcast networks use to better exploit audience labor: online broadcasting, audience-generated content, and social media as product placement. These strategies enhance the capacity of networks and advertisers to better measure audience labor through heightened surveillance available in new media. Understanding the surveillance and exploitation of audience labor by using these strategies is essential for empowering audiences as television and new media become increasingly integrated across platforms.

Keywords: new media, television, audience labor, surveillance

Introduction

Individualization of audience members is crucial to behavioral advertising, and television broadcast networks are pursuing a diversified portfolio of strategies to engage and surveil audience labor. This poster addresses the question: In what ways are television broadcast networks using new media tools to expand surveillance and exploitation of audience labor? The poster outlines three distinct strategies broadcast networks use to leverage new media to better surveil and exploit audience labor: providing online content, engaging audiences in new types of explicit labor or work through interactive or social networking tools, and normalizing the use of social media via program content. Outlining these strategies is a necessary first step for understanding the new political economy of audience labor when the lines between traditional broadcast television and new media are increasingly blurred.

Television Networks, Advertising and Audiences

Most television broadcast networks are based on the free-media model (Evans, 2008, p. 362-363). Central to the free-media model is the ability to surveil audiences. Traditionally, television networks would surveil their audiences by way of third party ratings services provided by marketing research companies, like A.C. Nielsen, that collect demographic and psychographic data through scientific sampling and intensive interview studies with audiences (Smythe, 2001, p. 258). Ratings are an important source of information for media and advertisers because they assure that the work of audiences is done in the most efficient way possible (Andrejevic, 2002, p. 236). Their control of the means of broadcasting enables the mass media and content providers to be brokers that sell the audience’s watching time to advertisers (Jhally and Livant, 1986, p. 130; see also Caraway, 2011).

Recent technological changes have fragmented audiences, changing the capacity of television networks to exploit and surveil audience labor (Turow, 2005, p. 111). Of greater concern are advances that allow audiences to by-pass advertising altogether, such as personal video recorders (PVRs). Meanwhile, the Internet provides new challenges and opportunities. The Internet appears to crowd out some people’s television watching (Liebowitz and Zentner, 2012). However, television programming is...
serialized, creating opportunities for time-shifting and online viewing with potential for expanding audiences rather than depleting them (Waldfogel, 2009, p. 161). It is this potential that broadcast television networks have embraced.

By encouraging audiences to participate in interactive Web-based media, networks are able to collect detailed information about audiences. Interactive websites expand and amplify the commodification process of audiences, particularly by “narrowcasting,” a strategy to increase the extraction of surplus value through extensive categorization and profiling (on narrowcasting in general, see Jhally and Livant, 1986, p. 139). Such narrowcasting facilitates personalized behavioral advertising, enabling networks to generate additional revenue and continuing brand loyalty (Andrejevic, 2002, p. 234-235; McStay, 2011, p. 319). By moving audiences onto websites, networks are also able to ensure engagement with content and minimize audience shirking. New media and online participation also serve to further deskill audience labor, making it as easy as possible to participate because networks directly measure audience labor through interactive features of programming or online content rather than cumbersome viewer surveys.

Online Advertising, Audiences & New Media

An informal content analysis suggests television networks use some combination of three different strategies for expanding their exploitation of audience labor using new media.

Online Broadcasting of Content

Television networks provide their broadcast content online, either on their own websites (e.g., ABC.com, Discovery.com) or third-party distribution websites (Netflix.com, Hulu.com). Some of these sites use display advertising, including placement of ads around the periphery of the screen or short video clips or commercials between content. Display advertising can be sold either through first-party (or direct) advertising, where the network or the media content provider (e.g. NBC or Hulu) collects viewer data and sells advertising space directly to advertisers, or through third-party (or indirect) advertising networks, where a third party, such as Google, in turn sells space to advertisers. Between 70 to even 95 percent of large content-providing websites (including traditional television and print media providers) use first-party direct sales (Evans, 2008, p. 385). While third-party advertising networks charge a fifteen to twenty percent commission on revenue (Evans, 2008, p. 385-386), they often have superior surveillance capabilities (McStay, 2011, p. 314-315). Third-party advertising networks also better tailor ads to viewers due to their larger pool of advertisers.

Table 1
Nexus of online programming considerations

<table>
<thead>
<tr>
<th>Online display advertising strategy</th>
<th>Site of program streaming or distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Network site (NBC, WB, etc.)</td>
</tr>
<tr>
<td>Direct (first-party) advertising</td>
<td>More direct oversight of audience labor, less detailed audience surveillance</td>
</tr>
<tr>
<td>Indirect (third-party) advertising (e.g., Google)</td>
<td>More direct oversight of audience labor, more detailed information from audience surveillance</td>
</tr>
</tbody>
</table>

Television networks face trade-offs about whether to provide their programming content online. Aside from the costs associated with developing sufficient infrastructure to stream programming on a network site, broadcast networks also must balance their desire to capture their existing audience to the
exclusion of other networks against their desire to poach or attract new audience members. Strong connections between particular programs and audiences may enable networks to force audiences to come to their site for content. Less popular shows may benefit from the exposure from appearing alongside other networks’ shows on a shared distribution platform, like Hulu.com. These considerations create a nexus of decisions (Table 1). The best strategy will vary according to the production costs (revenue needs) of the program or its existing audience size and loyalty. Networks may not be willing to pay for extra surveillance of audiences of particular types of shows, while shows with higher production costs will need larger audiences to attract the necessary revenue. Further, networks generate viewership for new offerings by predicting user preferences based on a detailed data profile.

**Audience-Generated Content**

Television networks expand audience labor through other forms of explicit labor, or work, and engagement via both online network-controlled websites and third-party fan or social networking sites. These outlets harness explicit work of audiences by outsourcing research and development about programming ideas to the audience and generating website content that becomes network property (Figure 1). These sites also act as unpaid focus groups, expand interest and participation in programming, and collect personal information about forum participants (Andrejevic, 2008).

**Figure 1.** Online viewer forum for *Dirty Jobs*. Source: http://community.discovery.com/eve/forums/a/ct/5141981108. Accessed September 24, 2012.

Similarly, networks engage audiences by creating interactive television shows based on voting or audience participation by phone or online. The audience labor, or participation, becomes part of the programming content. Since MTV pioneered audience participation in its show *TRL (Total Request Live)*,
popular competition shows, such as *Dancing with the Stars (DWTS)*, *American Idol*, and *The X Factor*, now encourage audiences to vote via toll-free phone numbers, SMS text messaging, online at the program’s website, or via Twitter. By suggesting particular folksonomies (e.g., #hashtags) during programming (Figure 2), networks are able to collect real-time viewer data, making the audience more valuable to advertisers. Such programs also discourage time-shifted viewing.

![Figure 2. Integration of social media during online broadcast of The X Factor. Source: Screenshot of The X Factor USA site (www.thexfactorusa.com) Twitter integration (24 September 2012)](image-url)

Television networks also exploit explicit audience labor, or work, through third-party fan or social networking sites. This type of audience participation, where users are granted pseudo-producer status, expands interest in and time spent watching future programming. Most importantly, it reinforces the link between programming, audience labor participation and surveillance. Viewer use of third-party social networking sites, such as Facebook and Twitter, also help advertise programming to new audiences, a form of targeted, word-of-mouth advertising to an audience member’s online social network. These innovations move audience labor online where it can be more efficiently engaged, surveilled, and sold to advertisers.

**Normalizing Social Media Via Program Content**

Networks also rely on product placement, or hybrid advertising, inside programming to capture audience labor. While product placement itself is not new, broadcast networks are now implicitly advertising new and social networking media, which serves to familiarize audiences with these tools and normalize their use, further facilitating audience labor (cf. Brown, 2006 on the normalization of surveillance in film). Examples include detectives using Facebook to solve crimes (*Psych*) and characters defriending each other on Facebook (*The Big Bang Theory*), as illustrated (Figure 3).
This integration of social networking websites into broadcast programming content is clearly a form of product placement, with the websites being the products being promoted by networks. The creation of “real” Twitter and Facebook accounts for characters (Figure 4) that mirror the story-lines and integration of the use of social networking media both online and offline, in broadcast content, provides added incentives for audiences to follow or interact with online network content. In these ways, networks are better able to surveil their audience and its labor and market their content to audience members’ online social networks.
Conclusion

The preceding theoretical discussion highlights the importance of the use of new media’s surveillance capacity for the effective capture of audience labor by television networks. By outlining three distinct new media strategies, the poster serves as a springboard for future research, which should focus on understanding when and why television networks use certain new media strategies and how the use of these strategies affects audiences’ enjoyment or perception of their television consumption. An understanding of these strategies, which now only resides among network executives, producers, writers, and advertisers who design new media content, is essential if audiences are to reclaim agency and some autonomy over their labor for networks.

References


Public Opinion Aggregation by Annotation and Tagging of Online News Stories

Pranay Sethi
School of Information, University of Michigan
sethlp@umich.edu

Keywords: sentiment analysis, public opinion mining, online commenting service, annotation, public opinion visualization

Abstract

Ubiquitous access to internet has resulted in more and more people going online to get their daily dose of news. In a 2010 survey conducted by the Pew Project for Excellence in Journalism, 41% of the respondents said they get most of their news online, 10% more than those who said they got most of their news from a newspaper. A lot of socio-technical factors have contributed to this phenomenal rise in adoption of online news in recent years. One of the biggest reasons why people are increasingly reading news online is because it facilitates discussion with peers (Nguyen 2010), offering different viewpoints which aid in forming a rounded personal opinion about the news story. The Pew survey found that 37% of online news users (and 51% of 18-29 year olds) think that commenting on news stories is an important feature to have. A lot of people tend to shape their opinion by reading discussion comments, reflective articles, blogs and even tweets about the news. Hence, an increasing number of people rely on online sources of news – be it news websites or news aggregator services like Digg, Reddit, Google Reader, Flipboard, Pulse etc.

The problem with these news websites and aggregators is that the only way people can gather public opinion is by actively searching through the endless stream of comments and feeds, filtering out spam (which is a growing problem) and then reading the relevant posts. A top trending story on Twitter will typically see multiple tweets per second, and keeping up with the rapid flow of incoming tweets is quite cumbersome and cognitively taxing. Hence it becomes increasingly difficult and time consuming for someone who wants to get the pulse of the people affected by a news story.

Furthermore, in certain scenarios people might want to look at more fine grained opinions. Currently, there is no elegant way to extract geographic and demographic impact of a news story. What is the public sentiment in Indonesia about the Arab Spring? How did the public opinion about the Wikileaks disclosures change as the story unfolded during the course of a year? It is very difficult and tedious to observe such patterns using the currently available news providers.

This work attempts to solve these problems by proposing a news aggregator platform which pulls news stories from various sources and also aggregates public responses, reflections, opinions and sentiments associated with those stories. This data is presented in ways that are easily understandable so readers can make better sense of the stories unfolding across the globe.

Such a news aggregator platform that gathers and display public opinion and sentiments about a story, must deal with various challenges –

1. Opinions are very subjective. Different people feel about a story in different ways. With such an enormous amount of diverse opinions and subjectivity, how can we possibly aggregate the responses into something that makes sense as a whole?

Acknowledgements: Prof. Yardi, S. for the guidance
Copyright is held by the author.
2. There isn’t really a unified web standard for expressing opinion (in textual form). Some people tweet in 140 characters, while others write elaborate blog posts. Some websites employ tags which a reader can use to define and classify their public opinion, while others rely on threaded comments and comment ranking systems. How can a platform be flexible enough to adapt to all these varied standards so that it can extract valuable data from various sources? Perhaps the platform can create a new standard of expression on the web which is flexible and comprehensive enough to be used to express diverse views about every news story in the world.

3. How to filter out spam while extracting public opinion?

4. Once the platform has access to the data it needs, how should it be displayed to the reader in a way that makes sense? What forms of visualizations, illustrations and graphical representations can be employed to give the reader a holistic view of how people feel about a story?

5. How can the platform determine and convey effects of geographical, demographic and temporal variations as the story unfolds?

These are just a few out of possibly many issues which must be dealt with. Previous research on similar public opinion aggregation services has greatly focused on natural language processing, data mining and text categorization and clustering. Xiaojun (2010) proposed a framework for crawling the web for comments and applying various data mining algorithms on the data to extract relevant information. Diakopoulos and Shamma (2010) used tweets posted in conjunction with the live presidential debate between Barack Obama and John McCain to gauge public opinion. Brody and Diakopoulos (2011) studied the use of word lengthening to detect sentiment in microblogs.

This research proposes a solution – The Opinionated Reader, which relies on sentiment tags and annotations associated with a news story. The essential idea is to create a commenting, discussion and sharing plug-in which can be used by news websites and aggregators as a commenting solution for their news pages. Users wanting to share or comment on a news story through the plug-in are asked to tag the news story with sentiment tags and annotate the story with their reaction (happy/positive or sad/negative). These tags and annotations are stored, aggregated and linked to each news story. A mobile application provides the front-end interface for users to access the news stories and the aggregated sentiment associated with each story. The basic architecture is explained as follows:

**The Opinionated Reader – Mobile/Tablet App**

The app fetches news articles from various web sources, based on the interests and preferences configured by the user. In every news article, a portion of the screen real estate is reserved for Opinions which shows graphical visualizations and illustrations of the public opinion surrounding the news story. These visualizations include:

- A Sentiment Graph indicating the popular tags associated with the story (E.g.: “Shocking”, “Inspiring”, “Amusing” etc.). See Figure 1 for example visualization.
- A Positivity Graph which plots the level of positivity associated with the story on a time scale from when the news broke. See Figure 2 for example visualization.

![Figure 1. Example Sentiment Graph](image1)

![Figure 2. Example Positivity Graph](image2)
The user can choose to see these visualizations for a particular time period in the evolution history of the news story, or for a specific country. The app also facilitates people to tag and annotate news articles from within its interface.

The Opinionated Reader – Commenting and Sharing Web Plug-in

These days, a common way of adding discussion and commenting functionality to news websites is by using 3rd party services (like DISQUS). The Opinionated Reader is a similar service which can be embedded into the news articles of various news websites to enable commenting and sharing. When someone wishes to comment on an article, the comment is directed through this plug-in, which allows the users to annotate the article with the sentiment tags and reactions along with their comments. The Opinionated Reader saves this information along with the commenter's location and date of comment (See Figure 3).

![Figure 3. The Opinionated Reader Web Plug-in - 'Add Comment' dialog mockup](image)

The Opinionated Reader – Back-end

The Back-end maintains a database of news items extracted from RSS feeds of various news websites. Each news article is linked with the sentiment tags and reaction/positivity annotations extracted from the comments and annotations gathered by the commenting plug-in. This data is used by the mobile/tablet app to generate visualizations (Sentiment Graph and Positivity Graph). The back-end also performs data mining on the tags and annotations for geographies and tracks the opinions across time.

The back end system responds to queries received from the mobile app with the news story and associated tags and annotations, which are then rendered by the mobile app for the user.
Discussion and Conclusion

This design idea is still in a nascent state and has long hours of research, brainstorming, designing and development to go before it can be realized into something tangible. Twitter has grown exponentially in importance as a news source and it would be vastly valuable to integrate Twitter with The Opinionated Reader. Possibilities include use of special hash tags and natural language processing of tweets to extract public sentiment.

The current design supports only two reaction annotations – positive and negative. Not every news story fits this annotation paradigm. Further research about human reactions to news stories might unveil interesting insights which would help zero in on a more robust annotation rubric.

Lastly, since this service is envisioned to be non-curated and non-moderated, the value served by the app depends on the users themselves. Greater adoption will lead to more annotations and tags, which translates into a more accurate public opinion as presented to the user.

References


Personal Information, Personal Property

Laura Lenhart
University of Arizona
lenhartl@email.arizona.edu

Keywords: personal information, information policy, intellectual property, privacy

Abstract

Personal information is valuable. Businesses and other organizations are able to draw on insights gleaned from personal information to develop useful goods and services and to pass on information that enables consumers to make better decisions. However, some collection, mining, and uses of personal information are ethically problematic. Target’s use of predictive analysis illustrates two problems posed by personal information: privacy and objectification. Target purchases and collects personal information to use in predictive analysis. The company has developed an algorithm to predict which customers are pregnant, in order to develop targeted marketing strategies (Duhigg, 2012). In addition to instituting private property rights in personal information, we need sphere-specific markets where information is exchanged. Cross-market exchanges should be blocked.

The problem of privacy is well-recognized and much discussed. Helen Nissenbaum has put forward an important and useful account of privacy arguing that privacy requires preservation of the norms that govern information exchanges. When we share information, we share it in specific contexts. Those contexts are governed by social norms that help shape our expectations about the flow of that information. A norm is a rule that governs our social relations and can be explicit or implicit (Nissenbaum, 2010). Personal information presents a threat to privacy because it is often collected, aggregated, and used in ways that do not respect norms of information flow (Nissenbaum, 2010).

The second problem is that some uses of personal information objectify and fail to respect persons. Aggregated data gives access to a very large range of a person’s activities and preferences and enables accurate predictions where data is not available. This information has market value in part because it is useful for influencing patterned behavior. Target is interested in knowing whether its customers are pregnant because shopping is mostly a product of habit rather than reasoned reflection. Most people do their shopping at many different stores and these habits are difficult to change. However, during periods of upheaval in our lives our habits are thrown off. Target wants to know which customers are pregnant in order to use targeted marketing to change their habits in a way that benefits the company (Duhigg, 2012). By offering coupons for items they may not associate with Target, the company pushes customers to pick up those items while picking the up items they do associate with the company. Very likely, these habits persist even after Target ceases to give coupons. Rather than thinking about customers as agents who use practical reason to make choices, this involves treating them like rats you are getting to go through a maze. When companies and other organizations try to leverage our unreflected, patterned behavior this way, they are treating us as objects to be manipulated by disregarding our own interests and attempting to structure our behavior in a way that benefits them (Nussbaum, 1995).

In order reap the benefits derived from the collection, mining, and exchange of personal information while reducing concerns about privacy, we should legislate private property rights in personal information that include:

- An exclusive right to sell the information in the relevant sphere.
- An exclusive right to transfer the information.
- A rights to prevent the collection of information--i.e., do not track.
- A right to destroy some information--to have some kinds of records expunged.
To discourage objectification, markets for personal information should be sphere-specific. Cross-market exchange should be blocked. Spheres are areas of social activity composed not only by specific norms, but united by a purpose and goal (Walzer, 1983). When information is tied to a particular sphere, it is less likely that we will see the individuals identified by personal information objectified. Their interests are reflected in the structure of the sphere itself. To illustrate, when the sphere of education functions well, the norms, values, and goals reflect the interests of students qua student as well as those of teachers qua teachers.

References

Information Systems and Smart Grid: New Directions for the IS Community

Jason Dedrick
Syracuse University
jdedrick@syr.edu

You Zheng
Syracuse University
yozheng@syr.edu

Abstract

While there is growing awareness that smart grid is an enabler of the new “green economy”, the IS academic community has been slow to recognize the challenges and opportunities it presents. So far there is little research in the IS field on smart grid technologies, yet this issue is closely linked to a number of common themes in the IS field. We identify some of the most salient issues in smart grid to the IS field and propose ways for the IS community to engage in this emerging example of the “internet of things”.

Keywords: smart grid, IS community

Introduction

The concept of smart grid is proposed as a potential solution to many current challenges in the electric utility industry: surging power demand, overburdened grid capacity and aging infrastructure. All these challenges make the grid vulnerable to frequent overloads and outages. The electric utility sector has also been criticized for its damage to the environment — the U.S. accounts for only 4% of the world population but accounts for nearly 25% of greenhouse gas emissions (U.S. Department of Energy, 2008) and the electric power sector is one of the three major emitters of greenhouse gases due to heavy use of coal.

In response to these concerns, smart grid has been identified as a potential solution to enable a more efficient and cleaner power supply by integrating various information systems and technologies into the current electrical platform (Illinois Smart Grid Initiative, 2008). Smart grid is not a single technology; rather it is a big and complex social-technical system that consists of many sub-systems and technologies. In general, there are four types of technologies in smart grid.

The first type is AMI (Advanced Metering Infrastructure) technologies that involve deployment of smart meters, customer systems, communication networks and back-office systems to enable two-way information flow between utilities and users (U.S. Department of Energy, 2012). A key benefit of AMI technologies is the ability to reduce peak demand through demand response programs and variable pricing. With AMI, utilities are able to monitor grid traffic in real time and take appropriate actions to reduce stress on the grid in peak hours, and users can also receive information such as electricity pricing on a real time basis and are better incentivized to manage and adjust their energy consumption (Morgan et al., 2009).

The second and third types of technologies involve new technologies in transmission and distribution systems that aim to improve system reliability and security. These technologies can extend the lifespan of existing infrastructure and avoid investments in new generation and transmission capacity. The last type of technologies includes those that extend the original scope of electricity grid, such as distributed generation, energy storage and plug-in electrical vehicles.

Current IS-Related Issues in Smart Grid

Smart grid technology has received some attention from academics, but mostly from an engineering perspective. So far there is little research in the IS field on smart grid technologies, yet there is potentially a very large area of smart grid that is closely linked to a number of common themes in the IS field. One of the many concerns is the interoperability problem (European Commission, 2009).
electricity grid has historically relied on proprietary technologies, but with smart grid upgrades, utilities are seeking to use new technologies from different vendors which raise the concern of interoperability. However, there are no widely accepted standards as of yet. Also, it is not clear whether the most advanced forms of smart grid technologies are compatible with the existing utility grid infrastructure. Without uniform technical standards, it will be difficult to achieve interoperability among different networks operating in many different regulatory jurisdictions (Zysman & Huberty, 2010).

Another concern is how to deal with big data (John, 2012). Smart grid systems will generate a huge amount of data that utilities have never faced before. For instance, a smart meter could receive and send energy usage information between utilities and customers every 15 minutes. Imagine a utility that has installed more than one million smart meters, each reporting in 96 times a day, and it is easy to see how much data could be involved. This will produce challenges for utilities that lack the systems and data analysis skills to deal with these data. This is another place where IS scholars can contribute.

Data security and privacy is another issue in smart grid (WestMonroe, 2012). Smart meters record detailed energy usage information in near-real time and such information would be attractive to third party companies (Munkittrick, 2012). Such likelihood would increase the tension between data access and data privacy. In general, without effective control, the abuse of customer information would reduce trust in both utilities and customers. Utilities are aware of this issue, but would benefit more from the expertise of the IS community.

Among the data generated by smart meters, how to capture right information and present it to customers is another interesting topic to IS researchers. Customers could receive their energy consumption information on a monthly, weekly and daily basis, and the content could also contain charts or graphs to present and compare the energy usage more vividly. Such a challenge not only requires good information visualization and presentation skills, but also involves a more important set of skills—data mining.

In addition to the aforementioned issues, the lack of IT skills poses a knowledge barrier to smart grid development. For utilities, smart grid requires new IT skills such as system engineering and system architecture that is beyond the reach of traditional power engineering. Thus, scholars with an interest in system development could fill this gap by focusing on topics like system development, HCI and system evaluation in smart grid. Also, IS economists have a role in understanding smart grid management. For instance, a transaction cost perspective could also be applied in smart grid to study whether some key technologies should be developed internally or purchased externally in the market.

**Future Directions**

As mentioned before, smart grid presents big research and educational opportunities for iSchools and the IS community. In this section, we would like to propose some general directions that would help the IS community to engage in this emerging, big topic. It is important to note that the following is not an exhaustive list of topics but represents some issues worthy of studying in IS field. Under each direction, we also propose some research questions that would help IS scholars better approach the smart grid.

1) **Information policy in smart grid:** focusing on policies for developing uniform technology standards, defining data ownership, and protecting consumer privacy. Some interesting questions include: What are the impacts of current information policies on smart grid adoption and management? What alternative policies or policy elements can governments adopt to encourage innovation?

2) **IS development in smart grid:** focusing on design and development of new applications such as new toolkits for customers at home to understand and manage energy or data management systems governing meter data. Particular interest could be emphasized on development methods, software and hardware development and user interface design. Sample research questions are: How could traditional system development theories in IS help to develop information systems in smart grid? How could user-interface design in technologies such as smart meters or home-energy-management products help to engage more customers in energy efficiency?

3) **Individual and organizational adoption and diffusion of smart grid technologies:** focusing on the determinants of adoption as well as impacts of smart grid adoption on different stakeholders. Research questions worthy of studying are: What are the determinations of smart grid adoption and what forces are required to overcome institutional obstacles to smart grid adoption in such a highly regulated industry? Are traditional innovation diffusion models adequate to study the adoption and diffusion of such large-scale complex system? If not, what new angles should IS scholars take to address the adoption and diffusion
question? How can actors with different visions, interests and motivations coalesce to achieve the collective action needed to reach widespread smart grid adoption?

4) IS use and management in smart grid: focusing on how certain information systems are managed in smart grid, learning models and processes during IT use in smart grid, and evaluation of system performance in smart grid. Research questions that might be of interest to IS researchers include: How do utilities deal with organizational challenges in smart grid implementation? How could knowledge and organizational learning perspectives help to address smart grid technology assimilation and implementation problems? To what extent does smart grid represent a discontinuous innovation that requires major organizational change? What are the impacts of smart grid technologies on utility performance?

5) Others: including other interesting topics such as customer education, smart grid curriculum development, and using social networks to build green customer community. Researchers interested in this track could focus on the following questions: How would IT and social networks help utilities to outreach customers in engaging in smart grid? How would smart grid curriculum be developed and fitted in the IS field?

Conclusion

Unlike many simple technologies that have been studied in IS field, smart grid is a big complicated system involving many different but interrelated subsystems. The complexity of the system and dynamics between its components provide many opportunities for IS scholars to study. However, current efforts in IS community are still rare. We hope the guidance in this study would attract more attention and interest from the IS field.

References


Ebooks and Cross Generational Perceived Privacy Issues

Michael Zimmer  
zimmerm@uwm.edu

Jennifer Thiele  
jennersthiele@yahoo.com

Renee Kapusniak  
kapusni2@uwm.edu

School of Information Studies, University of Wisconsin-Milwaukee

Abstract

Privacy issues, as it pertains to e-book downloads, are not well understood by the general user. People are downloading library e-books at an increasing rate, and participants of different age groups could potentially have different frameworks when conceptualizing online privacy. A common perception is that the younger computer user is less informed or concerned with privacy issues than the older user. However, recent studies have been finding that this is not necessarily the case. The researchers in this study would like to examine these privacy cross-generation perceptions when it comes to e-Book reading histories, utilization of commercial bookstore websites as well as library websites. Of particular interest are library e-books, as libraries are held at a different standard than commercial entities when it comes to user privacy. A small pilot case study will interview e-book users of different age groups to determine how they view privacy when purchasing e-books or when checking them out from the library.

Keywords: ebooks, privacy, older adults, young adults, libraries
Using Machine Learning Models To Interpret Disciplinary Styles of Metadiscourse in Dissertation Abstracts

Bradford Demarest  
Indiana University  
bdemares@indiana.edu

Cassidy Sugimoto  
Indiana University  
sugimoto@indiana.edu

Abstract

This paper presents the results of a study of disciplinary stylistic differences among dissertation abstracts from physics, psychology, and philosophy. Based on differences in relative frequencies of metadiscourse terms as provided by Hyland (2005), we used a machine learning approach to construct SMO vector support models of each discipline whose average accuracy (88.3%) surpassed a baseline model by 22%. We found that model term weights supported the findings of previous qualitative research regarding differences between disciplines and by extension between hard sciences, social sciences, and humanities. Given the success of the metadiscourse-based model, we conclude by proposing an expanded study to investigate disciplinary style both across disciplines and over time.

Keywords: metadiscourse, disciplinarity, machine learning, support vector model, dissertation abstracts

Introduction

Ken Hyland (2004) notes that academic disciplinary differences are not limited to topicality, but instead reflective of differences in “sanctioned social behaviours, epistemic beliefs, and institutional structures of academic communities” (Hyland, 2004, p.2). Discerning these differences in the writing of disciplines has thus far been mostly limited to qualitative or corpus methods, and has excluded machine learning based methods (a notable exception to this being Argamon, Dodick, and Chase's research (2008), which applied the SMO vector support model to investigate differences in epistemic language in between historical and experimental sciences).

Furthermore, research that has focused on disciplinary style as reflective of disciplinary beliefs and behaviors has entirely excluded dissertations, instead focusing on research articles. To address these two gaps in the research, the current study develops a machine learning based approach to investigate disciplinary style differences, using relative frequencies of metadiscourse terms in the dissertation abstracts of three disciplines: philosophy, psychology, and physics. These metadiscourse terms orient the author to the text itself as well as the reader in establishing epistemological and social norms. The current study's findings support the previous findings of qualitative and corpus-based studies (e.g. Becher, 1987; Hyland, 2008), which established epistemological and social differences among hard sciences, social sciences, and humanities.

Methods

The data used in this study was taken from abstracts for physics, psychology, and philosophy dissertations from the years 1980-1991 contained in the ProQuest dissertation database. Disciplines were operationalized by querying dissertations belonging to at least one subject category containing the string “physics”, “psychology”, or “philosophy”. Abstracts from dissertations with more than one identifying string (e.g. both “physics” and “philosophy”) were excluded from the dataset.

Acknowledgements: We would like to thank the National Science Foundation (Grant No. SMA-1208804) for financial support, as well as to ProQuest for sharing dissertation data with us for research purposes. Without their generosity this research could not be conducted.


Copyright is held by the authors.
Subsequently, data was divided into modeling data (taken from 1981, 1984, 1987, and 1990), development (from 1982, 1983, 1986, 1989, and 1991), and test data (from 1985 and 1988). Non-consecutive year groupings were chosen to create a model that could capture the evolution of disciplinary style over the decade. The training data set was then balanced for discipline frequencies; the discipline with the lowest number of abstracts was found (philosophy) and abstracts were randomly sampled from the other two disciplines until the training sample contained identical counts of all three disciplines. This yielded 4149 instances, or 1383 instances per discipline. Meanwhile, the test data set was collected as the set of all non-empty abstract records from 1985 and 1988, generating 11625 abstracts (874 of them philosophy, 7550 psychology, and 3201 physics).

For the set of features, a list of 316 words or phrases from six categories expressing interaction from Hyland (2005) was collected, and after removing 13 cross-category duplicates, the resulting 303 terms composed the feature set. These terms express authorial stance toward the text and engagement with the reader. Stance is expressed through hedges (which mitigate certainty), boosters (which amplify certainty), attitude markers (which express authorial affect), and self-mentions (with which the author alludes to herself), while engagement is expressed through imperative verbs and mentions of the reader (via pronouns or phrases like “the reader”). After collecting relative frequencies for the set of 303 terms, the WEKA machine-learning program (Hall et al., 2009) version 3.6.6 was used to create an SMO vector-support model (Platt, 1998) of each discipline in contrast to the other two disciplines (e.g. physics vs. non-physics). Each of these models was then tested against the test data set for classification accuracy.

Results

Table 1 presents the accuracy rate by percentage for each discipline, as well as averaged across all three models. The philosophy model was found to be the most accurate (with a 93.96% accuracy rate), and the psychology model the least (81.92%), but the average (88.3%) still outperformed a baseline classifier using the most-likely category by 22% (the most likely category for each model being non-discipline, generating an average 66% accuracy rate).

Table 1
Accuracy rates (%) for SMO Models

<table>
<thead>
<tr>
<th>Disciplinary Model</th>
<th>Accuracy (Percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physics</td>
<td>89.02</td>
</tr>
<tr>
<td>Psychology</td>
<td>81.92</td>
</tr>
<tr>
<td>Philosophy</td>
<td>93.96</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>88.3</td>
</tr>
</tbody>
</table>

Table 2 presents the features from the Hyland’s term set that were assigned absolute weights of 2 or more per discipline. In addition to the terms, the table also displays weights (positive valence indicating weighting in favor of a discipline, negative in favor of the non-discipline option in the model) and metadiscourse category to which the term belongs.

Discussion

More interpretation of these results is possible than space allows, but even a brief review reveals telling differences. The positively-weighted features that contribute most strongly to the SMO model of philosophy – “argue”, “thought”, “claim”, “think”, “know”, “my”, “establish”, and “true” suggest a field that is like Becher’s (1987) description of history: critical, reiterative, and “appealing to the professional judgment of the audience” (Becher, 1987, p. 273). Negatively weighted terms such as “observe”, “measure”, “increase”, “calculate”, and “use”, along with “we”, “known”, and “sure”, further support this depiction of philosophy, as these counter-terms imply non-philosophy as empirical, quantitative, and communal.
Table 2
SMO feature weights (absolute weight > 2)

<table>
<thead>
<tr>
<th>Weight</th>
<th>Term</th>
<th>Category</th>
<th>Weight</th>
<th>Term</th>
<th>Category</th>
<th>Weight</th>
<th>Term</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1922</td>
<td>argue</td>
<td>H</td>
<td>5.5801</td>
<td>assess</td>
<td>EM</td>
<td>-5.5028</td>
<td>assess</td>
<td>EM</td>
</tr>
<tr>
<td>4.5917</td>
<td>thought</td>
<td>B</td>
<td>3.8922</td>
<td>we</td>
<td>SM/EM</td>
<td>5.4482</td>
<td>observe</td>
<td>EM</td>
</tr>
<tr>
<td>4.392</td>
<td>claim</td>
<td>H</td>
<td>3.5686</td>
<td>recall</td>
<td>EM</td>
<td>4.7915</td>
<td>calculate</td>
<td>EM</td>
</tr>
<tr>
<td>-3.7582</td>
<td>observe</td>
<td>EM</td>
<td>3.3318</td>
<td>would</td>
<td>H</td>
<td>-4.3723</td>
<td>thought</td>
<td>B</td>
</tr>
<tr>
<td>3.6938</td>
<td>know</td>
<td>B</td>
<td>3.214</td>
<td>showed</td>
<td>B</td>
<td>-3.9784</td>
<td>argue</td>
<td>H</td>
</tr>
<tr>
<td>3.4845</td>
<td>my</td>
<td>SM</td>
<td>-3.2125</td>
<td>calculate</td>
<td>EM</td>
<td>3.9042</td>
<td>agree</td>
<td>AM</td>
</tr>
<tr>
<td>-3.4377</td>
<td>measure</td>
<td>EM</td>
<td>-3.0813</td>
<td>argue</td>
<td>H</td>
<td>-3.8457</td>
<td>think</td>
<td>B</td>
</tr>
<tr>
<td>3.3882</td>
<td>think</td>
<td>B</td>
<td>-3.0167</td>
<td>observe</td>
<td>EM</td>
<td>-3.7755</td>
<td>refer</td>
<td>EM</td>
</tr>
<tr>
<td>-3.0313</td>
<td>increase</td>
<td>EM</td>
<td>-2.7916</td>
<td>agree</td>
<td>AM</td>
<td>3.4943</td>
<td>known</td>
<td>B</td>
</tr>
<tr>
<td>-2.9822</td>
<td>we</td>
<td>SM</td>
<td>-2.4078</td>
<td>claim</td>
<td>H</td>
<td>-3.4344</td>
<td>my</td>
<td>SM</td>
</tr>
<tr>
<td>-2.6662</td>
<td>calculate</td>
<td>EM</td>
<td>2.3266</td>
<td>indicated</td>
<td>H</td>
<td>-3.386</td>
<td>know</td>
<td>B</td>
</tr>
<tr>
<td>2.5346</td>
<td>essential</td>
<td>AM</td>
<td>2.295</td>
<td>suggest</td>
<td>H</td>
<td>-3.3344</td>
<td>would</td>
<td>H</td>
</tr>
<tr>
<td>-2.5033</td>
<td>known</td>
<td>B</td>
<td>2.1503</td>
<td>find</td>
<td>B/EM</td>
<td>-3.0695</td>
<td>claim</td>
<td>H</td>
</tr>
<tr>
<td>2.3474</td>
<td>establish</td>
<td>B</td>
<td>2.139</td>
<td>appeared</td>
<td>H</td>
<td>-3.058</td>
<td>regard</td>
<td>EM</td>
</tr>
<tr>
<td>-2.2566</td>
<td>use</td>
<td>EM</td>
<td>2.1324</td>
<td>likely</td>
<td>H</td>
<td>-2.8161</td>
<td>key</td>
<td>EM</td>
</tr>
<tr>
<td>2.1499</td>
<td>true</td>
<td>B</td>
<td>2.1266</td>
<td>typical</td>
<td>H</td>
<td>-2.2775</td>
<td>disagree</td>
<td>AM</td>
</tr>
<tr>
<td>-2.0226</td>
<td>sure</td>
<td>B</td>
<td>2.0463</td>
<td>you</td>
<td>EM</td>
<td>-2.2689</td>
<td>recall</td>
<td>EM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-2.0432</td>
<td>show</td>
<td>EM</td>
<td>-2.2618</td>
<td>?</td>
<td>EM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.2183</td>
<td>indicated</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.2023</td>
<td>determine</td>
<td>EM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.1998</td>
<td>allow</td>
<td>EM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.1936</td>
<td>suggest</td>
<td>H</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.154</td>
<td>estimate</td>
<td>H/EM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.1313</td>
<td>essential</td>
<td>AM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2.0246</td>
<td>one's</td>
<td>EM</td>
</tr>
</tbody>
</table>

Note. Category Abbreviations: H = Hedge, B = Booster, EM = Engagement Marker, SM = Self Mention, AM = Attitude Marker.

Psychology is defined by its positive terms including “assess”, “we”, “recall”, “would”, “showed”, “indicated”, “suggest”, “find”, “appeared”, “likely”, “typical”, and “you” as a discipline that is communal, empirical (but, similarly to the way in which Becher (1987) describes sociology, self-conscious about knowledge’s status and methodology), while negative weighted terms (“calculate”, “argue”, “observe”, “agree”, “claim”) indicate the two extremes (one rhetorical and interpretative, the other objectivist and quantitative) between which psychology is positioned.

Physics is most strongly defined, based on positive weights, by “observe”, “calculate”, “determine”, “allow”, and “estimate” which along with other positive weighted terms (“agree”, “known”) suggest a discipline that is, in Becher’s words, “cumulative... tightly structured and atomistic” (Becher, 1987, p. 273), quantifiable and rife with directives for future researchers to expand upon the current work. Negatively weighted terms suggest a non-physics which is more interpretative (“assess”, “refer”, “indicated”, “suggest”) or persuasive (“claim”, “argue”, “think”, “disagree”).

Conclusion

Hyland’s terms serve as a useful feature set with which to model disciplinary voice, achieving reasonable levels of accuracy even when disproportionate distributions of classes exist between training and test data sets. Furthermore, the SMO machine-learning algorithm provides interpretable and insightful information at a term-specific level. That said, the current study has served as a useful pilot in that it has demonstrated a proof of concept. Further optimization of the algorithm and expansion of the feature set of terms to include synonymous terms could lead to even more accurate models, which we
propose to apply in a time-series analysis of major disciplines to analyze style shifts both within and across disciplines over the past century, based on the ProQuest dissertation data.

References


Extending the VIVO Ontology to iSchools: Enabling Networking of Information Scientists

Ying Ding, Erjia Yan, Ali Ghazinejad, Han Jia, & the VIVO ontology team
Indiana University

Abstract

VIVO, funded by NIH, utilizes Semantic Web technologies to model scientists and provides federated search to enhance the discovery of researchers and collaborators across disciplines and organizations. VIVO ontology is designed with the focus on modeling scientists, publications, resources, grants, locations, and services. VIVO data is annotated based on the VIVO ontology to semantically represent and integrate information about faculty research, teaching, and service. This paper introduces the birth and development of the VIVO ontology, and discusses the potential of expanding it to the information science community to facilitate networking of information scientists.

Keywords: VIVO, ontology, Semantic Web

Introduction

The exponential growth in complexity and scope of modern science has dramatically increased the demand for more collaboration among scientists in different fields and at different levels. Modern science is team-based, interdisciplinary and cross-institutional, but discovery across these boundaries is difficult. Researchers seeking answers to one research questions may have to consult scientists or studies from other scientific domains. Scientists have been heavily relying on the World Wide Web for supporting their research endeavors, especially for interdisciplinary and international collaboration (Hendler, 2003). However, currently popular Web technology is not satisfying for the needs of the collaborative and interdisciplinary “e-Science”. For example, boundaries of institutions, distributed data in different formats, and specialized terminology still impede the communication of scientific information between scientists. Therefore, new models of communication need to be forged so as to establish the next paradigm of tools of scientific collaboration on the Web.

The Semantic Web is designed to build a standard representation that can provide meaningful linkages across different sets of data to promote integration and communication. Many communities have embraced the Semantic Web technologies as a powerful and effective way to represent and relate data. The Linked Open Data (LOD) initiative currently contains 203 linked datasets which together serve 25 billion RDF triples to the Web and are interconnected by 395 million RDF links. The US government portal Data.gov makes around 400 of its datasets, summing to 6.4 billion triples, available as Linked Data. The Semantic Web technology may serve as an effective solution to the increasing and urging demand of broader and more in-depth communication between scientists in the academic community all over the world. However, there are currently no Semantic Web applications authorized to integrate official information of academic communities. VIVO can fill this gap.

VIVO is an open source Semantic Web application that, when populated with researcher interests, activities, and accomplishments, enables discovery of research and scholarship across disciplines and organizations. The VIVO core ontology models the academic community in order to

---

1 This poster is developed based on a poster presentation at the Conference on Semantics in Healthcare and Life Sciences (CSHALS2012).

2 http://www.data.gov/semantic/index


Copyright is held by the authors.
provide an consistent and connected perspective on the research community to various shareholders, including students, administrative and service officials, prospective faculty, donors, funding agencies, and the public (Karfft, Cappadona, Devare, et al., 2010). The major impetus for NIH to fund the VIVO effort to “develop, enhance, or extend infrastructure for connecting people and resources to facilitate national discovery of individuals and of scientific resources by scientists and students to encourage interdisciplinary collaboration and scientific exchange”.

VIVO can support discovering potential collaborators with complementary expertise or skills, suggesting appropriate courses, programs, and faculty members according to students’ interests, and facilitate research currency, maintenance and communication. The implementation of the VIVO ontology to iSchools allows one to search across experts among varied sub-domains in the field of information studies. In this paper, we present a relatively comprehensive discussion of the development of the VIVO core ontology.

VIVO Ontology Development

According to Gruber (1993), an ontology is a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. Simply put, it contains a hierarchical taxonomy or controlled vocabulary and secondary associations between terms. Complex real world entities can be expressed by assigning properties (i.e, relationships/associations) to classes/subclasses. Subclasses usually can inherit properties from their upper classes. For example, faculty member have subclasses assistant professor and associate professor. If faculty member has the property teach class, then assistant professor and associate professor will also have the property teach class. Furthermore, ontologies also contain inference rules to enable machine-processable computing and reasoning (Berners-Lee, Hendler, & Lassila, 2001). An ontology may express the rule that if woman A is the mother of woman B, and woman B is the mother of man C, then woman A is the grandmother of man C. With the reasoning power, machines can manipulate the concepts in a well-defined logic way that can be easily understood by human beings (Guarino, 1998; Guarino & Giaretta, 1995). An ontology models the semantics of components by defining concepts (classes) and the relationships (properties) between them. The VIVO ontology is a unified, formal, and explicit specification of information about researchers, organizations, activities and relationships that link them together.

Modeling Principles

A significant challenge in developing the VIVO core ontology is to include enough detail to allow for meaningful cross-site discovery of data, while keeping it simple enough to apply to diverse academic and clinical institutions (Karfft, Cappadona, Devare, et al., 2010). The development of the VIVO ontology follows four principles:

- Core and localization: The design of VIVO core and localization combines the advantages of standardization with those of customization.
- Modularized design: It subdivides a system into independent but connected modules that can be used in different contexts to enable multiple functionalities.
- Aligning with other ontologies: VIVO ontology is iteratively refined to align with other ontologies.
- Modeling complex relationship as classes: In VIVO ontology, complex relationships are modeled as entities in order to add properties to these relationships.

Conceptualization

The major classes covers the three important areas of the faculty research activities are: research (bibo:Document, vivo:Grant, vivo:Project, vivo:Software, vivo:Dataset, vivo:ResearchLaboratory), teaching (vivo:TeacherRole, vivo:AdvisingRelationship), and services (vivo:Service, vivo:CoreLaboratory, vivo:MemberRole). For each of the three parts, there are existing ontologies that conceptualize them.

The VIVO core ontology, on one hand, reuses the useful classes and properties of FOAF and, on another hand, develops customized properties and classes targeting at modeling social network of scientists. For information resources, BIBO ontology models publications; DOAP models projects; and

---

Doblin core models documents in general. Other resources, including courses and grants, are usually organized and stored in databases. The VIVO core ontology reuses useful classes and properties from those existing ontologies, and further connects, creates, or modifies those distributed classes in a uniform framework of modeling social network of scientists.

Figure 1 shows concept map of VIVO’s (version 1.3) main classes and objective properties. Bubbles in different colors represent different sources of the classes. Lines between bubbles indicate their relationships, which are object properties in VIVO ontology.

![Figure 1 VIVO’s main classes and objective properties](http://sourceforge.net/apps/mediawiki/vivo/nfs/project/v/vi/vivo/8/83/VIVOCoreOverview.1.3.byPropertyGroup.2011-08.png)

**Interoperability**

A key issue in building the VIVO ontology is enabling interoperability among different ontologies. Since ontologies have been developed in different contexts, ontology interoperability is a key factor essential for aligning and integrating distributed ontological resources over the Internet. Mappings to these ontologies will enable VIVO data to be shared among a variety of systems. The development of the VIVO ontology reuses several commonly used ontologies, including upper level ontologies Event ontology, Basic Formal Ontology-OBO Foundry), as well as domain ontologies (i.e., FOAF, Geopolitical ontology, SKOS, and BIBO).

**Localization**

Individual institutions can localize, or extend, the ontology to support local requirements. Different namespaces are adopted for different localized versions in specific institutions. Individual installations may extend the core with ontologies that reflect available data sources according to their contextual needs. While all institutional installations of VIVO share the core ontology, each institution is free to

---

4 High definition figure can be found at: [http://sourceforge.net/apps/mediawiki/vivo/nfs/project/v/vi/vivo/8/83/VIVOCoreOverview.1.3.byPropertyGroup.2011-08.png](http://sourceforge.net/apps/mediawiki/vivo/nfs/project/v/vi/vivo/8/83/VIVOCoreOverview.1.3.byPropertyGroup.2011-08.png)
extend this ontology or add additional ontologies as desired. The VIVO core ontology plays the role as an integration layer that permits data from different institutions to be queried in a consistent way.

Conclusion and future development

The **iSchools** Project is a consortium of 36 institutions located at 11 countries\(^5\); meanwhile, the studies of information is interdisciplinary. VIVO provides a practical ontology that represents **iSchools**’ needs by providing local value for sustainability, while putting those institutions who adopt it at a competitive advantage for discovery through linked data and for compliance with federal data initiatives. In the future, we intend to expend the VIVO ontology to all **iSchools** and expect to see an enriched knowledge base of academic communities among **iSchools** and even beyond. There are at least three ways to participate in VIVO: 1) Download, Adopt, and Implement: The open source VIVO software and ontology are now available for download. 2) Provide Data: **iSchools** can participate by providing machine readable data for research discovery. Bibliometric and funding data are of great interest to the research community. 3) Develop Applications: Many software applications can benefit from using information that will be provided by the national network of VIVO. New applications can use information from the national network to provide enhanced search, new collaboration capabilities, grouping, finding and mapping scientists and their work\(^6\).

References


\(^5\) [http://www.ischools.org/site.descriptions/](http://www.ischools.org/site.descriptions/)

\(^6\) [http://vivoweb.org/participate](http://vivoweb.org/participate)
The It Girls Overnight Retreat: An iSchool Recruitment Model for Prospective Female Undergraduate Students

Dori Farah  
Syracuse University  
School of Information Studies  
dfarah@syr.edu

Julie Walas  
Syracuse University  
School of Information Studies  
jlwalas@syr.edu

Abstract

Responding to the well-known need for more women in computing and information science, the Syracuse iSchool offers a powerful and replicable solution: The It Girls Overnight Retreat. This intensive, peer-focused program combines mentoring with a specifically designed curriculum to engage, inspire, and celebrate young women and their potential in the information industry. First piloted in 2011, the iSchool received college applications from 50% of the seniors in attendance and 11 of them became matriculated students. The event model was then enhanced with research-based assessment tools, activities for parents and high school counselors, and the inclusion of barrier-breaking female role models. In Fall 2012, the second-annual event attracted 93 high school juniors and seniors from 43 high schools across 6 states. For the iConference, Syracuse will report on recruitment results, pre- and post-survey data results from the attendees, and best practices for peer institutions with a desire to replicate the event.

Keywords: women in computing, innovation in educational practices, STEM recruitment tools, best practices, female undergraduate recruitment

Introduction

In the United States in 2009, women earned 57% of all undergraduate degrees, but only 18% of all computer and information sciences undergraduate degrees (National Center for Women in Information Technology, 2012). A lack of gender diversity in undergraduate computer and information science programs results in a shortage of female technology talent entering the job market, reduced innovation and competitiveness in the field (Page, 2007), and decreased customer satisfaction with new technologies and products (Catalyst, 2004).

Approach

Our Academic Alliance with the National Center for Women in Information Technology (NCWIT) has provided access to proven practices for recruiting women into information technology (Barker, Cohoon, & Sanders, 2010). We have infused NCWIT’s research and best practices into every aspect of The It Girls Overnight Retreat. For example:

Scholarship in Action

Research says that young people express a desire to do something socially relevant and interactive. The It Girls Overnight Retreat includes workshops on the use of technology for social change and culminates with an “Overnight Challenge” – a project completed in small groups of peers and mentors that is focused around the concept of Scholarship in Action, challenging them to create innovative solutions for the world’s problems by leveraging technology.

Acknowledgement: Special thanks to Katie Hassman, Ph.D. student at the Syracuse University iSchool, for her hard work on the assessment model and continuous support of The It Girls.


Copyright is held by the authors.
Influencing the Influencers

NCWIT stresses the importance of tailoring your message to the influencers – parents and counselors. The It Girls Overnight Retreat agenda includes parent- and counselor-specific sessions focused on career trajectories in I.T., research-supported best practices for encouraging women in I.T., the college application and financial aid processes, and the flexibility and marketability that an education in computer and information science provides.

Leveraging Assets

Leveraging your assets, such as the University, alumni, community members, and current students is crucial in recruitment. The It Girls Overnight Retreat is strategically timed during a typical high school holiday weekend, Homecoming events with alumni, and a busy University-wide Undergraduate Admissions Visit Day. Because of this, attendees have the opportunity to interact with mentors and role models of all ages and backgrounds, the environment is upbeat and exciting, and stay overnight in the residence halls with current female iSchool students.

Results

In total, 43 of the 93 attendees of the 2012 It Girls Overnight Retreat completed the pre- and post-event surveys. The following descriptive statistics are based on survey data from those 43 participants:

- 85% indicated that the retreat greatly increased their awareness of women in technology.
- 30% indicated that they found technology more interesting than they did prior to the retreat.
- Prior to the retreat, 63% of participants indicated that they would not consider a career in technology. Of these 63% of participants, 41% indicated that they would consider a career in technology after attending the retreat.
- Prior to the retreat, 28% of the respondents indicated that it was highly likely they would obtain a career in technology. After the retreat, 52% of respondents indicated that it was highly likely they would obtain a career in technology.
- Prior to the retreat, 37% of respondents indicated that they planned to apply to the iSchool. After the retreat, 49% indicated that they planned to apply to the iSchool.

Conclusion

The event was deemed a success for the Syracuse iSchool in 2011, so we are repeating the event in November 2012. We are currently working to improve the event model for dissemination and will be implementing researched-based surveys to properly assess the event model and support our findings. At the 2013 iConference, we anticipate reporting the following discoveries from the pre- and post-event data collected from attendees: a positive change in their interest in I.T. as an educational focus and career; an increased confidence in their own technical abilities; a heightened capacity to visualize a future career for themselves in I.T.; and a stronger desire to study in a technical program and pursue a technical career.

Implications

The It Girls Overnight Retreat brought positive recruitment results to the Syracuse iSchool, but its impact could be more significantly realized if replicated at other iSchools. We are excited about the positive growth in gender diversity this event could bring to our iSchool community and the difference it could make in our world.
References


The Globalization of the iSchools Movement

Laurie J. Bonnici  
University of Alabama  
lbonnici@slis.ua.edu

Heidi Julien  
University of Alabama  
hjulien@slis.ua.edu

Kathleen Burnett  
Florida State University  
Kathleen.Burnett@cci.fsu.edu

Abstract

In 2005 a relatively small interdisciplinary group of LIS schools, all based in the U.S., announced its intention to form a new “iField.” The explicitly stated goal behind the formation and formalization of this group was the coming to grips with the “elusive identity [that] poses a challenge for the I-School movement” (King, 2006). Today 40% of the iSchools Caucus is non-U.S. based. This research examines the impact of the international member schools on what was once an exclusively American group. The internationalization phenomenon is examined from the perspective of the Information Outcome Space (Gross & Latham, 2011). Content analysis of school websites addressing vision and mission statements, “about the school” statements, and messages from the Deans/Directors were conducted to discern the philosophical approaches of iSchool as they relate to the concept of information. This research addresses whether information conception is the uniting, identifying, and defining identity for the iCaucus.

Keywords: iSchools, Library and Information Science, Computer and Information Science, content analysis, international

Introduction

Over the past few decades, shifts in the professional marketplace, globalization, and a rapidly changing technological landscape have complicated the disciplinary identity formation process of technology-intensive disciplines. The disciplinary identity of LIS has been contested since its origins in 19th century librarianship training programs (Burnett & Bonnici, 2006). Inter-professional and interdepartmental competition, jurisdictional disputes—including LIS and computer science over the emergence of information technology as a discipline—have problematized the establishment of a lasting disciplinary identity (Bonnici, Subramaniam, & Burnett, 2009). Conversations among visionaries from both disciplines resulted in the formation of the iSchools Caucus in 2005. In an effort to determine whether this i-movement is a nascent discipline or repackaging of old ideas under a new guise, research has been carried out on three aspects of iSchools scholarship: integration (organizational culture and leadership), discovery (research and teaching), and application (professional associations and practice).

Integration

The term information science was coined in the early 1960s, and applied almost immediately to the description of two very different fields. The two disciplines in which Information Science most often resides, Library and Information Science (LIS) and Computer and Information Science (CIS), historically claimed distinctly separate domains. While the titles associated with the two disciplines have persisted, variations have been tested and discarded. Informatics, suggested by Gorn as an alternative to Computer and Information Science, had short-lived currency in each of the disciplines. This term survives, however it failed to rebrand the disciplines on a mass scale. Many of the academic units that house LIS programs removed “Library” from their names, but their programs continue to focus, albeit to varying degrees, on libraries as institutions and librarianship as practice.
Discovery

The removal of “library” from the disciplinary identity of LIS schools is a result of the influence of information science, impacting the “and” in Library and Information Science. The impact of information systems on LIS as a discipline has notably been an influential factor in the iSchools movement, though not the main factor (Bonnici & Burnett, unpublished). The most notable distinction between the CIS and LIS disciplines is research domain. CIS research has its roots in information systems while LIS research has centered on the human element in information processes, widely known as information behavior research. Saracevic (1999) declared that information science falls short of being a full-fledged discipline due to the lack of cross-disciplinary connection between systems-centered and user-centered research engagement. Evidence indicates that communication between CIS and LIS researchers has begun to blur the lines of disciplinary boundary. Despite the disconnect, Saracevic noted progress indicating that information retrieval research has recognized motivational or affective relevance as elements in text retrieved by a system. Has the merging of the disciplines through the iSchools movement found common ground in the information sciences of LIS and CIS?

In 2005 a relatively small interdisciplinary group of LIS schools, all based in the U.S., announced its intention to form a new “iField.” The explicitly stated goal behind the formation and formalization of this group was the coming to grips with the “elusive identity [that] poses a challenge for the I-School movement” (King, 2006). The iSchools Caucus created the term iField to capture this elusive identity, and defined it as:

an academic field of study and a professional career field that deals with all the issues, opportunities, and challenges we face in our emerging Information Age…. The iField addresses this fundamental issue: how do we harness that incredible flow of information for the betterment of society, rather than get swamped by it? (iSchools Caucus, n.d.)

Today the iSchools Caucus is made up of thirty-eight member schools with nearly 40% of the membership being non-U.S. based.

As viewed through the theoretical lens of the Chaos of Disciplines (Abbott, 2001), as applied by Bonnici and Burnett, the culture of the iSchools Caucus has disciplinary breadth, is self-replicating in method, and has progressed through a method of rediscovery. The mechanism of progression from LIS to iField is an inverted fractal cycle, moving from specific disciplinary focus to broad focus over time. The inverted directional move is contrary to the theoretical view that disciplines move from broad concerns to more specific foci over time. Preliminary findings indicate that the move from the specific domain of libraries toward a more encompassing concern with information more generally may have fueled the inverted fractal cycle.

Bonnici and Burnett’s research spanned 1965-2009. Since that time iCaucus membership has grown, with the most notable trend in membership being the addition of schools outside of the U.S. This research project examines the impact of these new schools on what was once an exclusively American group of schools.

Research Questions

RQ1: How has internationalization impacted the iSchools movement?
RQ2a: How do U.S. iSchools and International iSchools differ in their conceptualization of information?
RQ2b: How is information conceptualized by the iCaucus?
RQ3: What possible effect might internationalization of the iSchool movement have on the future development of information science as a part of CIS, LIS, and/or a new iField?

Theoretical Framework

Interviews with the Deans and Directors of the six founding iSchools identified three unifying themes; people, information, and technology. In order to gain insight on the inverted fractal cycle, we will examine the internationalization phenomenon, from the perspective of the Information Outcome Space proposed by Gross & Latham (2011). The model consists of three tiers, with information conception
dominating the top tier. Information conception is defined by focus on information as product over process. More precisely stated, information conception is the outcome of the search for information rather than concern with approaches to finding information. However, information seeking is conceptualized at the mid-level tier. Theoretical concepts at the middle tier include information people, information technology, and information quality, key factors in the information seeking process. These three concepts were identified by the founding iSchool deans when asked to describe the factors uniting the iSchools’ membership. The six administrative heads stated “people, information, and technology” as the key elements uniting the schools under the iCaucus label. Information technology conception is characterized as the use of technology as a primary way to disseminate and acquire information. Information people conveys the idea that people are considered valuable as information resources, whether they deliver information or facilitate the finding of information. The third concept at the middle tier of the model is information quality. Quality refers to information that is tangible and useful. These three elements alone do not equal information conception. Rather, they are key elements in the process of finding useful information. However, these three elements currently unite and thus define the iSchools movement. We have a notion that information conception (information as product) embodies additional unidentified factors. These factors may be found in the combination of LIS and CIS disciplines that constitute the iField, explaining why even the iSchools Caucus considers its identity “elusive.”

**Methods**

Content analysis of school names and sections of school websites (accessed September - December, 2012) for the six founding iSchools and the international schools were examined. Aspects for analysis included vision and mission statements, “about the school” statements, and messages from the Deans/Directors were conducted to discern the philosophical approaches and attitudes of iSchool members as they relate to the concept of information. Content analyses are phase one of a two-phase project that will include interviews with iSchool Deans/Directors internationally, following the conference. Phase two analysis will also include an examination of faculty disciplinary contribution through a review of educational and career backgrounds.

**Analysis and Conclusions**

Data analysis was conducted in terms of the middle tier of the Information Outcome Space. In addition, analysis was informed by invisible substrates (Bates, 1999). Bates’ “Three Big Questions;” physical, social, and design aspects of information were used to determine information conception. The researchers attempted to determine if the non-U.S. member schools contribute unique views to the iSchools group as it relates to information people, information technology, and information quality.

Data analysis addressed the logical, methodological, and pragmatic relations among and between the two disciplinary areas that populate iSchool membership: CIS and LIS. Our goal in this research is to facilitate understanding of the future trajectories of the discipline of Information Science, including the impacts of globalization of the information disciplines. Data analysis sought to determine if international CIS-based versus LIS-based member schools differ in their conception of information conception. The overarching purpose of this research is to determine whether information conception is the uniting, identifying, and defining identity for the iCaucus.

**References**


Community Informatics Studio: A Model of Information Scholarship in Action

Martin Wolske mwolske@illinois.edu
Colin Rhinesmith crhines@illinois.edu
Jennie Archer enger2@illinois.edu
Emily Bayci bayci1@illinois.edu
Ryne Leuzinger leuzing2@illinois.edu
Lucas McKeever lmckeev2@illinois.edu

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Abstract

This poster describes an educational research project we call “Community Informatics Studio,” which uses studio-based learning (SBL) to support enculturation into the field of community informatics. The SBL approach is rooted in the apprenticeship model of learning in which students study with master designers or artists to learn their craft and is closely related to John Dewey's inquiry-based learning. The poster reports on research from our Community Informatics Studio course, which uses experiential learning as a model of information scholarship in action. We highlight the guiding theoretical frameworks, research questions, methodological approaches, findings, and recommendations. The goal of our poster is to respond to this year’s conference theme by presenting our research question: “How can the Community Informatics Studio be understood as an innovative educational model of information scholarship in action?”

Keywords: community informatics, studio-based learning, digital literacy, popular education, service-learning

Introduction

This poster presents findings from an educational research project we call “Community Informatics Studio,” which uses studio-based learning (SBL) to support enculturation into the field of community informatics (Keeble & Loader, 2001; Gurstein, 2003; Williams & Durrance, 2009). The SBL pedagogy is rooted in the apprenticeship model of learning in which students study with master designers or artists to learn their craft and is closely related to John Dewey's inquiry-based learning (Lackney, 1999). The poster reports on research from our Community Informatics Studio course, at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, which uses experiential learning as a model of information scholarship in action. We highlight the guiding theoretical frameworks, research questions, methodological approaches, findings, and recommendations. The goal of our poster is to respond to this year’s conference theme by presenting our research question: “How can the Community Informatics Studio be understood as an innovative educational model of information scholarship in action?”

Theoretical Framework

The relevant literature in which our research is situated can be traced to the teachings of John Dewey, Paulo Freire, and other educational philosophers who argued that learning must be rooted in people’s everyday experiences (Becker, 2000; Mehra, 2004; Ball, 2008). “Service-learning” seeks to connect “practical experience to reflection” as a strategy to enhance student learning and “personal growth through civic engagement” (Ball, 2008). However, scholars have noted the long-standing tension that exists between theory and practice in library education (Morehead, 1980). Others have sought to...
problematize the concept of service-learning by addressing the unequal power relationships in university/community partnerships (Reardon, 1998; Mehra, 2004). Our educational research project attempts to fill a gap in the literature by fusing theory and practice through studio-based learning (Lackney, 1999; Brown, 2006; Brocato, 2009) as an innovative pedagogical approach to advance library and information science.

**Case Study: Community Informatics Studio**

Using studio-based learning (SBL) methods, Community Informatics Studio has brought students, instructors, and professionals together with community members in a collaborative environment to apply community informatics scholarship to a real-world problem or “case.” Project work has been student-led, with students’ works-in-progress, critiques by the instructors and peers, and engagement with outside practitioners all taking place inside the studio: a facility designed to closely reflect the professional work environment. The learning objective is to shift focus from “learning-about” to “learning-to-be” by providing a more nuanced understanding of the thinking behind design choices (Brown, 2006). Formal lectures are used sparingly to provide students with the background information and knowledge needed to successfully address the case and produce an effective proposal/plan.

**A Model of Information Scholarship in Action**

In this section we briefly describe our research this semester. We emphasize the use of SBL as an experiential approach to promoting information scholarship in action. **Fall 2012.** Students have built upon lessons learned from the past two semesters: the case for the course in 2010 was titled “Designing Public Computing Centers of the Future” and the case in 2011 was titled “Community Media Newsrooms: Equipping Citizen Journalists.” This semester, students are addressing a new case, “Popular Technology” (Eubanks, 2011). Eubanks’s concept of popular technology is rooted in popular education, participatory research, and participatory design methods. Students are engaged in SBL with the goal of developing more engaged digital and media literacy workshops. The poster will describe findings from a “participatory evaluation” (Stoecker, 2013, p. 199) of the students’ work with instructors by analyzing the following three data sets: weekly journal entries, desk-critique artifacts, and final presentation materials. The workshops developed by students for the course will be made available online at www.prairienet.org and offered to community members at public computing centers as part of our “Eliminate the Digital Divide” grant, awarded to two of this abstract’s authors by the Illinois Department of Commerce and Economic Opportunity.

**Conclusion**

In this poster we hope to introduce an innovative educational model of information scholarship in action by presenting our case study of the Community Informatics Studio course, which uses studio-based learning to advance library and information science. Our goal is to inspire further lines of inquiry to support the development of engaged information scholarship.

**References**


Nonparametric Estimation of Search Query Patterns

Soohyung Joo  
sjoo@uwm.edu  
Dietmar Wolfram  
dwolfram@uwm.edu  
Suyong Song  
songs@uwm.edu

University of Wisconsin-Milwaukee

Abstract

In this poster, we adopted nonparametric regression as a method to identify the unique distribution of query log data collected from the Excite search service in May 2001. In Informetrics, parametric modeling has been widely used in tracing term frequency data, such as Zipf's law, Lotka's law, or Bradford's law. However, these traditional parametric methods have had limited application when detecting distributions for large datasets with a nonlinear pattern and a long tail. This study tested kernel regression as an alternative tool to model nonlinearity of term frequency patterns. The results indicated that the kernel regression produced an improved model fit compared to previous parametric approaches in modeling query patterns.

Keywords: power law, non-parametric estimation, kernel regression, query log analysis

Introduction

In informetrics, many researchers have attempted to generalize the patterns of different types of information production. Based on observed regularities in the distribution of data, mathematical models are developed to match the observed pattern (Wolfram, 2003). Query logs have been a focus of research to informetricians for several decades, and it is widely known that many size-frequency patterns show a reverse-J-shape with a long tail. This reverse J-shaped distribution has been modeled using parametric methods, in particular power law curves (Newman, 2005). Parametric models identify a specific equation, and attempt to estimate corresponding parameters from the observed data. For example, Zipf’s law, which posits two parameters in its power law equation, has been most widely applied to explain a relation between term frequency and its rank. However, modeling based on power law sometimes misspecifies nonlinear patterns due to the noise mainly caused from long tails. Also, as the size of a dataset grows, it becomes more difficult to adequately fit the observed frequency distributions to mathematical models (Ajiferuke, Wolfram, & Famoye, 2006).

Nonparametric regression can be a compelling alternative to predominant power low method. Nonparametric regression is more flexible and accurate to articulate the nonlinearity of any function, as it determines the local shape of the conditional mean relationship (Blundell & Duncan, 1998). This study intends to adopt nonparametric estimation, to be more specific–kernel regression--as a way to model search query patterns. To the best of our knowledge, this study is the first attempt to apply kernel regression to estimate term frequency patterns.

Kernel Regression

The kernel regression is one of nonparametric methods in statistics that estimate a non-linear relation between random variables. The kernel regression can be represented with a regression function, g(x):

\[ y_i = g(x_i) + \epsilon, \quad i = 1, 2, \ldots, n \]  


Copyright is held by the authors.
where \( g(x) \) is the unknown regression function and \( \varepsilon_i \)'s are the independent and identically distributed zero mean errors (Eubank, 1999).

Unlike parametric regression, \( g(x) \) is a form of unknown smooth function. In this case, \( g(x) \) can be estimated based on the following nonparametric regression function:

\[
\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \tag{2}
\]

where \( K(\cdot) \) is the kernel function (weight function) that penalizes distance from the local position where the approximation is centered. \( h \) is the bandwidth that controls the width of the weight function (Takeda et al., 2007).

The selection of the kernel and bandwidth determines the shape and the smoothness of the estimation in kernel regression. The selection of the kernel function is open. \( K(\cdot) \) can be selected from a Gaussian, Epanechnikov, or other forms that satisfy the requirements of the kernel function:

\[
\int K(z)dz = 1; \int zK(z)dz = 0; \int z^2K(z)dz = c \tag{3}
\]

where \( c \) is a constant value (Takeda et al., 2007). In this study, an Epanechnikov kernel was selected as a weight function, which is commonly used in kernel regression:

\[
K(u) = \frac{3}{4}(1-u^2)1_{[|u|\leq1]} \tag{4}
\]

Silverman’s (1986) rule was employed for the bandwidth calculation (where \( \sigma \) is standard deviation of \( x \) and \( n \) is sample size):

\[
h = 1.06 \times \sigma \times n^{-\frac{1}{5}} \tag{5}
\]

Data Collection and Analysis

Search term frequency data extracted from the 2001 Excite query transaction log data set (Spink et al., 2002). The dataset consists of 587,145 non-repeating queries submitted to the Excite search engine in May 2001. Individual terms were identified within queries by parsing for standard delimiters, with exceptions for entries such as URLs and email addresses, which were treated as single terms. The term frequency distribution represents the number of terms that occur one time, two times, etc., up to a maximum of 129,170 times. The dataset yielded 1,538,120 tokens for 182,012 term types. The tokens represent the individual occurrences of specific terms. The term types represent the distinct terms entered by searchers. Even though the dataset is not very recent, it was appropriate to test the new method as it has shown typical query patterns represented by power law.

As is usually done in power law function fits, both the size of the query term (x-axis) and the number of query terms that occur with a given frequency (y-axis) were transformed using logarithmic scaling. This study fitted the transformed data using both parametric and nonparametric methods. In parametric estimation, both linear modeling and polynomial modeling, in particular quadratic equation modeling, were examined. In nonparametric analysis, kernel regression was conducted using MATLAB. Root Mean Squared Error (RMSE) and \( R^2 \) values were used to compare the model fits across three different estimations.
Results

First, we modeled the size-frequency function of query terms using both parametric and nonparametric methods. The observed data were fitted to test the power law identified in Equation 6, essentially a Lotka function:

$$Y = A/X^b \quad (6)$$

where $A$ and $b$ are parameters to be estimated. To easily visualize the patterns, we applied a logarithmic transformation to both axes:

$$\ln(Y) = \ln(A) - b \ln(X) \quad (7)$$

$$y_i = c + b_1 x_i \quad (8)$$

where $\ln(Y) = y$, $\ln(A) = c$, and $\ln(X) = x$. In addition, quadratic regression was employed to achieve a better model fit:

$$y_i = c + b_1 x_i + b_2 x_i^2 \quad (9)$$

To the same function, we applied kernel regression based on an Epanechnikov kernel. As shown in Table 1, the kernel regression resulted in the lowest RMSE ($0.154$, $R^2 = 0.956$), while linear and quadratic models resulted in values of .312 and .195, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>0.312</td>
<td>.821</td>
<td>3.763 -1.270</td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.195</td>
<td>.930</td>
<td>5.994 -3.275 .426</td>
</tr>
<tr>
<td>Kernel</td>
<td>0.154</td>
<td>.956</td>
<td>(non-parametric)</td>
</tr>
</tbody>
</table>

Using the parameters estimated from ordinary least squares (OLS), we plotted the obtained linear and second-order polynomial curves. As shown in Figure 1, the linear equation had some defects in correctly articulating the pattern of actual observations. In particular, there were evident limitations in modeling low frequency terms and the long tail. The curve was slanted toward the x-axis due to the relatively long tail. The y-intercepts of the linear and quadratic (polynomial) models were 3.763 and 5.994 respectively, far from the observed value (5.050).

On the other hand, the result of the kernel regression showed an improved estimation result by achieving a better RMSE value. The kernel regression result not only specified the unique feature of the long tails but also precisely followed the nonlinear pattern in the region of highly ranked terms.

Second, we modeled the token distribution by term frequency using the same methods. As the pattern was shaped as “V”, a linear model was not appropriate in this case ($RMSE=6.263$; $R^2=0.172$). Quadratic model showed a better model fit ($RMSE=0.195$; $R^2=0.676$) as it depicts “U” shape. Kernel regression exhibited a better model fit by achieving an RMSE value of 0.154 ($R^2 = 0.798$).
Figure 1. Size-frequency distribution model fits (log transformed)

Table 2 Estimation of token frequency distribution (log transformed)

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>6.263</td>
<td>.172</td>
<td>3.763 - .270</td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.195</td>
<td>.676</td>
<td>5.994 -2.275 .426</td>
</tr>
<tr>
<td>Kernel</td>
<td>0.154</td>
<td>.798</td>
<td>(non-parametric)</td>
</tr>
</tbody>
</table>

Figure 2 compares three lines of different estimation methods. As shown in Figure 2, the kernel regression presented more elaborated estimation than the other two parametric regressions.
Conclusion

This study introduced nonparametric estimation to model query term frequency distribution in Informetrics. We compared first- and second-order regression models and a nonparametric method. This study supports that kernel regression could be useful in exploring query distributions, or potentially other frequency distribution data for large, nonlinear datasets with long tails. Effective modeling is particularly important for accurately predicting outcomes for informetric data. This has applications for more accurate estimation of informetric model outcomes not only in information retrieval environments, but also more generally for information production and use.

Classical parametric methods, represented by a power law in this example, rely on a specific model and seek to compute associated parameters in the presence of noise mostly coming from the long tail end and nonlinearity. In contrast to parametric methods, nonparametric estimation relies on the observed data itself to dictate the structure of the model (Wand & Jones, 1995).

In this poster, we described and proposed the use of kernel regression as an alternative tool for tracing query term frequency distribution patterns. This study yields some methodological contributions for the field of informetrics. Nonparametric methods can be applied to different types of data such as size-frequency form distributions used in journal productivity, author productivity, citation distributions and indexing exhaustivity distributions. As many data distributions found in metrics studies exhibit non-linear patterns and long tails, nonparametric regression can be useful to model observed data, particularly for larger, more difficult to model datasets which are now common.
References


HELPP Zone: Towards Protecting College Students from Dating Violence

Amirreza Masoumzadeh¹, Lei Jin¹, James Joshi¹, Rose Constantino²
¹School of Information Sciences, ²School of Nursing, University of Pittsburgh
[amm215, lej17, jjoshi, rco100] @ pitt.edu

Abstract

Dating violence is a type of Intimate Partner Violence (IPV) which is a serious, preventable public health and social justice issue. If dating violence is left unchecked, it can bring short term and long term negative effects to survivors, especially college students who are developing emotionally. We built a mobile application called HELPP Zone (Health, Education on safety, and Legal Participant Preferred) for college students to protect them against potential dating violence. In the HELPP Zone, a user can set a list of trusted contacts and her schedule, and reach them for help based on the context (user’s location, schedule and status) when potential violence occurs. The HELLP Zone is our first step towards creating a platform to end dating violence.

Keywords: dating violence, protection, mobile application

Introduction

Intimate partner violence (IPV) is a serious, preventable public health problem. It is described as physical, sexual, or psychological harm inflicted by a current or former partner or spouse (CDC 2011). Dating violence is a type of IPV and has brought serious public health and social justice issues to people, especially the young who may not report it because they are afraid to tell friends and family (Black and others 2010; CDC 2012). If the dating violence is left unchecked, it will bring short term and/or long term negative effects to the survivors (Catallozzi and others 2011). Young people are developing physically, emotionally, and socially are heavily influenced by their relationship experiences. Unhealthy, abusive or violent relationships have negative proximate and long term consequences (Claiborne 2012; Garcia and others 2012). For example, survivors of dating violence are more likely to have poor grades, school attendance, and disruptive behaviors. They may binge drink, attempt suicide, and argue and fight with friends and family.

There have been few mobile apps developed recently with the goal of preventing violence. Circle of 6 (2011) allows a user to choose a group consisting of six trusted contacts, and provides quick communication with them if needed. OnWatch (2011) is a similar solution with few additional pay-only features. However, in both of these apps, the group of trusted friends is static and cannot adapt to the various needs of a user in different situations. For example, if a student is in the university area, she may include classmates, who may not be useful if she needs help in a different geographic area as on a spring break or vacationing at home with family. Furthermore, a user may not choose to include her parents as trusted contacts in every situation. She might prefer to ask for help from a close friend or sibling in difficult dating situations. We believe that such context-focused dynamic contact grouping is necessary in order to be in touch with the most appropriate contact or resource which the HELPP Zone app provides.

We built a mobile app, called HELPP Zone (Health, Education on safety, and Legal Participant Preferred) to prevent potential and escalating dating violence. In building the HELPP Zone App, we addressed the fundamental paradigm of Diffusion of Disruptive Innovations (DDI) (Fisher and Clayton 2012). A DDI is defined as a deliberately and intentionally developed product that functions as a disruptive innovative tool. The HELPP Zone app was designed as a simple product that enters the transformative service arena as a common activity. It offers an inexpensive and convenient service that has the potential to be diffused as a disruptive tool in preventing or stopping relationship violence among
college students. It provides several ways of disrupting an ongoing or escalating dating or intimate partner violence through just-in-time communication and intervention from and by appropriate and trusted contacts. In HELPP Zone, a user selects a dynamic list of trusted contacts, which are called helpers. Helpers can be reached by text messaging or phone calls based on the user's geographic location, time of day and need. In addition, the app is equipped with information (e.g. safety tips) that are ticker tape-like cues or messages to detect a controlling date/partner, to exit from a beginning violent episode or de-escalate a violent situation (Constantino and others 2007). The HELPP Zone app is free and currently available for Android as an inexpensive and widespread platform (HELPP 2012). We will make it available on other popular platforms such as iOS in the near future.

**HELPP Zone Features**

In the followings, we describe the settings and safety features available in the HELPP Zone app. These features are demonstrated in Figure 1.

**User Settings**

*Status.* In HELPP Zone, a user can set her status (or statuses), which represent her current situation, e.g., walking, biking, dating, in a restaurant, in a mall, etc. The user's status can be automatically updated based on her event/location.

*Event.* A user can define a series of events in the location tab, shown in in Figure 1.B and 1.C. Each event determines the user's status, the time when it will be active, the period that the event lasts, and as an option, the location where the event is taking place. An event can be either one time (e.g., a specific date) or periodic (walking or biking).

*Helpers.* A user can choose contacts that she can store in the mobile phone as helpers. These are user-trusted contacts that are known by her as individuals who will come to her rescue unconditionally without hesitation or being judgmental (Figure 1.D and 1.E). The helpers are either active or inactive based on the user's current status. During emergencies, a user can press the "Call 911" button at the "Action" screen, to quickly dial 911 and be transferred to the call screen of the phone.

**Safety Feature: Calling Emergency Numbers**

*Call 911.* During emergencies, a user can press the "Call 911" button at the "Action" screen, to quickly dial 911 and be transferred to the call screen of the phone.

*Call Hotlines.* By pressing the "Hotlines" button at the "Action" screen, a user is given the choice of calling one of the essential national hotlines.

*Call Campus Emergency.* College students would be more efficiently served if they call campus police for unwanted or reportable events in the campus area, or other safety-related offices in a specific university as shown in Figure 1.F and 1.G. A user may call emergency numbers that are available in the application database, such as campus police, sexual assault services based on her choice of a university in the settings page.

**Safety Feature: Situation-Aware Helpers**

Different set of helpers may be trusted and contacted for different situations. Helpers become active depending on the user's current situation which consists of user's current status(es), location and timestamp. When the user selects a helper, she could indicate in what status the helper should be reached (Figure 1.E). When it is in the time span of an event and a user is in the corresponding location (if specified), the user's status is updated to the specified status based on the pre-defined user's schedule. Such dynamic update to the user's status(es) provides situation-awareness to the application which is needed for selecting appropriate helpers.

**Safety Feature: Asking for Help**

When there is at least one active helper based on a user's current situation, the user can reach active helpers through the following ways:
Call Me. Suppose the user is in a situation where she needs to stop an unwanted or unconsented activity in a non-noticeable way and an incoming call to her phone can give her a chance to disrupt the activity and let her helpers know of any potential threat to her. By pressing the “Call Me” button, a prepared message, as shown in Figure 1.H, will be drafted to be sent to the helpers that reads: “HELPP: Please give me a call. I need to pretend I am talking to you”.

Need Help Here. When a user needs to let her helpers know about her location, she can use this function: Need Help Here as shown in Figure 1.I. A prepared message, which contains the location coordinates of the user and current street address, will be drafted to be sent to the helper(s).

Ask Something. When a user needs to consult her helpers, a prepared message will be drafted to be sent to the helpers by pressing the button: Ask Something

Safety Feature: Educational Resources

Multiple educational resources are available to the users by tapping the Resources icon on the “Action” screen.

On-Screen Ticker Tape Tips. As shown in Figure 1.A, HELPP Zone includes a push-like feature that shows quick safety messages on a message bar at the bottom of the “Action” screen. These messages are educational and informational but short and non-intrusive. A different message is displayed each time the user opens the application or goes to the home tab. The user has the option of circling through the messages by tapping on the message bar.

Safety Instructions. We provide concise safety instructions for intimate partner violence survivors, including protective tips when the user is at/away from home, including tips on rights, reporting, legal and law enforcement processes. As shown in Figure 1.J, a user can press the “Resources” button to check out these instructions.

Figure 1. HELPP Zone’s features
Conclusions

In this paper, we introduced a mobile application, called HELPP Zone, for college students to prevent or disrupt potential dating violence. Technology is here to stay. It can be used as a sword or as a shield (IOM 2012). Our next step is to examine the acceptability and feasibility of the HELPP Zone app among college students through evidence-based research and respond to these questions: Would someone use this app? In what way will users of HELPP Zone become engaged and feel rewarded? What are the risks in disrupting ongoing dating violence by text messaging and/or calling helpers/emergency numbers? Will the educational materials (tips and instructions) and resources help in identifying beginning signs of abusive and coercive dating behaviors? Will the ease and period-of-time of app use inform the user to visit only safe and popular dating venues that are data-based on the crime rates in the location of and the ratings of the venues?

References

HELPP Zone, Retrieved from www.sis.pitt.edu/~leijin/HELPP.html
OnWatch, Retrieved from www.onwatchoncampus.com
Academic Genealogy as an Indicator of Interdisciplinarity: A Preliminary Examination of Sociology Doctoral Dissertations

Chaoqun Ni
Indiana University Bloomington
chni@indiana.edu

Cassidy R. Sugimoto
Indiana University Bloomington
sugimoto@indiana.edu

Abstract

This project aims to explore the interdisciplinarity heritage of the field of Sociology by utilizing academic genealogy networks based on doctoral dissertation data from ProQuest database. With the advisor name provided by each dissertation, the discipline of each advisor was identified by matching back to ProQuest database. The Subject Categories assigned to advisor dissertations were approximated as the advisor discipline. Results provide an indication of the major areas interacting with Sociology. Disciplines such as business administration, health science and economics show increasing trends in the percentage of sociology advisors getting degrees from them, while the decreasing trends can be found in history and literature.

Keywords: Sociology, interdisciplinarity, academic genealogy, dissertation, ProQuest

Introduction & Literature Review

This project utilizes academic genealogy as an indicator of interdisciplinary for sociology. Academic genealogy is a method that involves creating networks of scholars by connecting doctoral advisors with their students. This can be used as a tool for studying the interdisciplinarity of a field by examining the degree to which advisors are imported from other disciplines (Sugimoto, Ni, Russell, & Bychowski, 2011). Journal articles, authors, and citation relations between these have often been used in scientometrics to identify interdisciplinary relationship (e.g., Leydesdorff, 2007; Porter & Rafols, 2009; Rafols & Meyer, 2010; Schummer, 2004). However, bibliographic publication based approaches of interdisciplinarity study suffers from certain limitations. Firstly, the impact of some disciplines with average citation rates will be overestimated. Secondly, if based on collaboration relationships, many contributions are neglected: Ni, Sugimoto and Jiang (to appear) found that about 70% of library and information science journal publications have a single author. Finally, it is not be easy to measure the interdisciplinarity of those disciplines publishing in genres other than journal articles.

Doctoral dissertations provide a useful alternative for scientometric research. All research disciplines produce dissertations; therefore, this genre does not favor certain disciplines. Each individual produces only a single dissertation in each discipline; therefore, dissertations are not skewed in the direction of subdomains or authors who might be inordinately prolific. Finally, dissertations provide the opportunity to study mentoring through advisorship, a relatively unexplored network structure for scientometric research (Russell & Sugimoto, 2009).

Data Collecting & Processing

This project relies on the data provided by the ProQuest dissertation database. The data set (hereafter PQuest) was provided from ProQuest and covers about 2.3 million dissertations from 1,490
research institutions across 66 countries from 1848 to 2011. For more detailed information about the database, please refer to (Ni & Sugimoto, 2012).

In this project, Sociology dissertations are operationalized as dissertations assigned with the PQuest Subject Category (SC) Sociology and sub-categories. As shown in Figure 1, a single dissertation may be assigned to multiple SCs (e.g., 0635-Sociology, Organization Theory; 0451-Psychology, Social; 0350-Health Sciences, Education, where the first four digits is a unique ID for each SC). In such a case, the dissertation is considered as a Sociology dissertation, but could easily have been completed in another field. Therefore, this operationalization should be considered as a mere proxy for disciplinarity.

The major task in this project was to identify the advisors of Sociology dissertations and the disciplines in which the advisors completed their degrees. In PQuest, the full name of each dissertation author was provided (e.g. “GREGORY, EDWARD WYLIE, JR.”). However, the advisor information is not available in many cases: only about 70% of Sociology dissertations contain advisorship information. Those that did contain advisorship information often had incomplete data (e.g., “D. ADAMCHAK” instead of “DONALD J. ADAMCHAK”). Therefore, the first stage was to correctly identify advisors for each Sociology dissertation and then match them with their dissertation. Advisor-author name matching and disambiguation was performed using the following steps:

- Author set: the original author full name of all dissertations in PQuest was split into first name, last name, middle name, nick name and suffix (if any);
- Advisor set: the original advisor name of each Sociology dissertation was split into first name, last name, middle name, nick name and suffix (if any). [The majority of advisor names only have the last name and first name initial];
• In the advisor set, advisors of two dissertations were identified as the same one if: the
dissertations they mentored were from the same university, and their last name, first name initial,
middle name initial and suffix are the same. After this step, about 80% of advisors were matched.
• Match advisor set with author set using different strategies.

It should be noted that some dissertations provide committee members in the “advisor” field. This
project parsed all the names provided and took the person appeared first as the advisor of that
dissertation. Here we do realize that the name matching and disambiguation of advisor information is by
no means perfectly accurate.

Results & Analysis

Description of Sociology in PQuest

The earliest Sociology dissertation in PQuest was completed in 1894. As of 2011, there are
96,434 dissertations assigned with this SC. Table 1 shows the number of dissertations in Sociology by
decade, as well as the percentage of dissertations with advisor information.

Table 1 Dissertations and advisor-available dissertations in sociology by decade

<table>
<thead>
<tr>
<th>Decade</th>
<th>#Dissertation</th>
<th>%Advisor-available dissertation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1900's</td>
<td>1</td>
<td>0.00%</td>
</tr>
<tr>
<td>1900's</td>
<td>5</td>
<td>0.00%</td>
</tr>
<tr>
<td>1910's</td>
<td>4</td>
<td>0.00%</td>
</tr>
<tr>
<td>1920's</td>
<td>1</td>
<td>0.00%</td>
</tr>
<tr>
<td>1930's</td>
<td>38</td>
<td>5.26%</td>
</tr>
<tr>
<td>1940's</td>
<td>53</td>
<td>3.77%</td>
</tr>
<tr>
<td>1950's</td>
<td>939</td>
<td>1.63%</td>
</tr>
<tr>
<td>1960's</td>
<td>2,667</td>
<td>0.52%</td>
</tr>
<tr>
<td>1970's</td>
<td>7,862</td>
<td>0.50%</td>
</tr>
<tr>
<td>1980's</td>
<td>10,896</td>
<td>26.52%</td>
</tr>
<tr>
<td>1990's</td>
<td>35,359</td>
<td>84.74%</td>
</tr>
<tr>
<td>2000's</td>
<td>32,494</td>
<td>92.86%</td>
</tr>
<tr>
<td>2010-2011</td>
<td>6,115</td>
<td>93.97%</td>
</tr>
<tr>
<td>Total</td>
<td>96,434</td>
<td>71.39%</td>
</tr>
</tbody>
</table>

Each of the 13 sub-categories is considered specialties of Sociology. Table 2 shows the number
of dissertations in each specialty, ordered by the year each specialty first appeared in the database.
Table 2. Number of dissertations in each specialty and year of first appearance

<table>
<thead>
<tr>
<th>Specialty</th>
<th>#Dissertation</th>
<th>Year_First_Appear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sociology, Social Structure and Development</td>
<td>11882</td>
<td>1894</td>
</tr>
<tr>
<td>Sociology, Ethnic and Racial Studies</td>
<td>21073</td>
<td>1900</td>
</tr>
<tr>
<td>Sociology, Public and Social Welfare</td>
<td>9600</td>
<td>1912</td>
</tr>
<tr>
<td>Sociology, Industrial and Labor Relations</td>
<td>7602</td>
<td>1914</td>
</tr>
<tr>
<td>Sociology, Theory and Methods</td>
<td>5065</td>
<td>1922</td>
</tr>
<tr>
<td>Sociology, General</td>
<td>12768</td>
<td>1930</td>
</tr>
<tr>
<td>Sociology, Criminology and Penology</td>
<td>11325</td>
<td>1934</td>
</tr>
<tr>
<td>Sociology, Individual and Family Studies</td>
<td>23395</td>
<td>1934</td>
</tr>
<tr>
<td>Sociology, Demography</td>
<td>4108</td>
<td>1935</td>
</tr>
<tr>
<td>Sociology, Organizational</td>
<td>2494</td>
<td>1964</td>
</tr>
<tr>
<td>Sociology, Organization Theory</td>
<td>417</td>
<td>1971</td>
</tr>
<tr>
<td>Sociology, Environmental Justice</td>
<td>67</td>
<td>2008</td>
</tr>
<tr>
<td>Sociology, Sociolinguistics</td>
<td>166</td>
<td>2008</td>
</tr>
</tbody>
</table>

Dissertation Advisor Institutions & Countries

Sociology dissertations in PQuest are from 678 institutions across 26 countries, with 92.63% of total dissertations coming from institutions located in United States (US), which is probably due to the fact that PQuest mainly focuses on North America.

For those 68,845 dissertations with advisor names, 62,072 (90.16%) of them were matched with corresponding advisor dissertations, representing 26,208 unique advisors. Some of those advisors mentored several doctoral students: one advisor mentored 68 students, another 52. There are 72 advisors who mentored more than 20 doctoral students in PQuest. Those 26,208 unique advisors graduated from 522 institutions across 15 countries. Result shows that 23,898 advisors received their degrees in the US, and 2,005 from Canada. The University of Wisconsin - Madison is the largest exporter of Sociology advisors, followed by University of Michigan and Columbia University. Table 3 provides the top 10 institutions by the number of sociology advisors they exported.
Table 3 Top 10 Exporting Institutions of Sociology Advisers

<table>
<thead>
<tr>
<th>Institution</th>
<th>#Sociology Advisors</th>
</tr>
</thead>
<tbody>
<tr>
<td>The University of Wisconsin - Madison</td>
<td>840</td>
</tr>
<tr>
<td>University of Michigan</td>
<td>822</td>
</tr>
<tr>
<td>Columbia University</td>
<td>779</td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>691</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>588</td>
</tr>
<tr>
<td>Stanford University</td>
<td>583</td>
</tr>
<tr>
<td>University of Illinois at Urbana-Champaign</td>
<td>556</td>
</tr>
<tr>
<td>The Ohio State University</td>
<td>520</td>
</tr>
<tr>
<td>Yale University</td>
<td>470</td>
</tr>
<tr>
<td>University of California, Los Angeles</td>
<td>435</td>
</tr>
</tbody>
</table>

Disciplines

The discipline of advisors is approximated using the SC that is assigned to the advisor’s dissertation. Therefore, to measure the interdisciplinarity of Sociology via academic genealogy, this project identifies the disciplines in which advisors received their degrees. In PQQuest, some dissertations have multiple SCs. Of all the 26,208 dissertations by Sociology advisers, 21,241 (81.05%) were assigned to a single discipline.

There are 166 disciplines in PQQuest. Sociology advisors were assigned to 90 different disciplines. Table 4 shows the top 10 disciplines that appear most frequently. The results indicate that Education and Psychology are large contributors of Sociology advisors. It perhaps indicates that many of advisor dissertations were actually completed in these fields, with emphases in Sociology.
The percentage of advisers from each discipline changes over time. Figure 2 displays the percentage of advisers from sociology by decade, and figure 3 displays the percentage of the rest nine of the top 10 disciplines by decade. It shows that Education is one of the major disciplines where sociology advisers got their degrees, and the percentage by each decade is not stable. The increasing trends of sociology advisers getting degrees from business administration, health science and economics can be found, while the decreasing trends can be found in history and literature. Due to space limitation, further analysis on the change of other discipline percentages, as well as single and multiple discipline dissertations will be provided later.
Several limitations were encountered in this project that should be remedied in future work. First, not all dissertations have advisor information and those that do are fairly recent. For better diachronic studies, manual data collection of these advisors will be necessary. Second, author-name disambiguation is still not perfectly accurate and requires additional refinement. Lastly, using SCs as a proxy for disciplinarity introduces some limits to interpretations, as it is unclear which of the dissertations were actually conducted in departments of sociology.

This poster provides the first large-scale of academic genealogy using PQuest and automated matching algorithms. Future work will attempt to refine these methods and add additional disciplines in order to generate a better understanding of the interaction of disciplines through academic genealogy. Additionally, future work should seek to merge this data source with others for an increased understanding of disciplinary mobility and productivity. Three categories of analyzing the disciplines that have great impact on the

References


Dispatches, Digests and Doodles:  
Exploring the Significant Properties of Field Notebooks

Andrea K. Thomer  
thomer2@illinois.edu

Nicholas M. Weber  
nmweber@illinois.edu

Center for Informatics Research in Science and Scholarship  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign

Abstract

Whether a first-person narrative, a structured observational data source, or simply a diary of doodles, field notebooks are the material instantiation of most site-based social and natural sciences. In this poster, we present preliminary work that explores the significant properties of field notebooks across three diverse disciplines: Paleontology, Archaeology, and Field Biology. This work is meant to supplement existing norms in digitization processes, and begin moving many of those projects towards a more general, interoperable approach to preserving the rich content held within field note collections.

Keywords: digital preservation, field notes, natural history, museum informatics, digitization.

Introduction

Beyond the silos of the LAMs: making field data broadly accessible

Field notes – the recording and cataloguing of observations in site-based field-work – often contain a huge amount of data structure in their writing, yet libraries, archives and museums typically do not have the resources or domain expertise to reveal that structure. Creating a typology of this structure and making it machine-readable will only become more important as more notebooks are not just digitized but also transcribed. Currently efforts to make that structure discoverable through annotation or text markup are nascent-to-nonexistent (Thomer et al., 2012).

What’s significant about significant properties?

Much has been written about the preservation of digitized cultural objects from a curatorial standpoint, including a differentiation between representation properties, which describe the data content of a digital object, and significant properties, which describe the informational content of a digital object (CEDARS, 2002). Our use of significant properties in this poster is similar to Knight and Pennock, and Wilson’s definition: “...significant properties are defined as the characteristics of an information object that must be maintained to ensure that object’s continued access, use, and meaning over time as it is moved to new technologies” (2011, p.163; 2007). Here however, we want to emphasize the importance of maintaining an intelligible transfer of the content from a recorded observation found in a field notebook to the digital instantiation of that content, which may be viewed as both a surrogate of the original analogue text and a data point (for a more thorough discussion see Thomer et al., 2012).

Methods. We performed a comparative analysis of individual field notebooks from three different scholarly disciplines (palaeontology, archaeology, field biology), and reviewed domain-specific literature describing best practices to help ensure that our analysis was commensurate with field practices of that discipline in a general sense, as opposed to field practices specific to a particular person or expedition.

Acknowledgement: Many thanks to Carolyn Sheffield at the Smithsonian’s Field Book Project (http://www.mnh.si.edu/rc/fieldbooks/) for providing us with digitized notebooks.


Copyright is held by the authors.
In assessing each field notebook for its significant properties, we followed Grace and Knight’s typology of significant properties: content, context, rendering, structure and behavior (2008). Below we map those significant property types onto questions we answered in analyzing the field note collections from our three chosen disciplines:

- What temporal division is most frequently recorded, and how (e.g. day, field campaign, year)? **Context**
- Is there a cataloguing or pagination scheme in place? **Structure**
- How are external sources of data included or cited (e.g. sketched, pasted in, cited)? Are external databases referenced? **Content**
- What structured data is collected and how is that structure represented? **Rendering**
- What and how is quantitative data collected? **Behavior**
- What events are recorded? **Content**

**Analysis**

**Paleontology (Simpson, 1930; Whitmore, 1975; Walcott, 1879).** In paleontology, field books are labeled with expedition name, date range, book number within a larger set (e.g. “Book 1”), and author name.

![Field Note Book](http://research.amnh.org/paleontology/notebooks/simpson-1930a/01.jpg)

*Figure 1. Index to George G. Simpson’s notebook, 1930. Available from [http://research.amnh.org/paleontology/notebooks/simpson-1930a/01.jpg](http://research.amnh.org/paleontology/notebooks/simpson-1930a/01.jpg)*

Entries are divided by day, but the books are divided by both subject and campaign (that is, each expedition is recorded in a new series of notebooks). Pages are hand numbered and are often continuous from one book to another within a series (e.g. in Simpson’s case, book two starts on page 167). The inside cover of the first book in a series includes a list of specimen numbers that are referenced in the book (this list would be added after the book was ‘completed’), as well as an index to other field books describing other aspects of the expedition (Figure 1) – particularly the geologic context surrounding each collected specimen -- thus creating a kind of database-of-notebooks all related to one particular
expedition or site. While not every paleontologist keeps this same system of notebooks, many have similar practices; “locality” ledgers containing detailed stratigraphic and geologic data are common (if not necessary).

![Figure 2](http://bit.ly/TQ1YOe)

Field Biology (Grinnell, 1911; Rafinesque, 1818). Many biologists write in their notebooks in a manner that is similar to, but less geologically-oriented than, paleontologists: many maintain a pagination scheme that stretches over many notebooks (possibly over the duration of their entire careers), and they mix personal narrative with semi-structured descriptions and lists of the animals and plants observed while travelling. Biologists conducting field work often follow a variation of the “Grinnell Method” of data collection: plant and animal names are underlined; their number or even absence are noted; and collected specimens are recorded in a table including a catalog number, their gender, a map reference, and additional notes describing their appearance or behavior (Perrine & Patton, 2011). Grinnell himself did not seem one to sketch (though many other biologists do sketch their subjects), but he did on occasion paste annotated maps into the front or back pages of his journals detailing where, exactly, he did his work.

Archaeology (Strong, 1933; Smith, 1938). Archaeological field books tend to be temporally bound by the length of an expedition, and the ‘entries’ are recorded in frequency by day. In some of the books we surveyed, we noted a high amount of marginalia; one author (Strong) seems to have summarized some of his narrative into short parenthetical statements to the left of the main text. Archaeological journal entries are also less systematically laid out than in other disciplines; the author often relies on the pre-printed page numbers in the upper corners of each page instead of creating his own numbering system, and if artifacts were collected, their catalog numbers are not as prominently or systematically recorded.
While these notebooks do contain data in the form of maps and detailed drawings of pottery sherds, these diagrams are primarily used to augment the author’s daily narrative describing his travels, meals, and conversations with other people. More extensive and structured data collection often happens outside the field notebook and in the context of formal excavations.

Figure 3. A page from Strong’s notebook. [http://www.nmnh.si.edu/naa/features/strong2z.htm](http://www.nmnh.si.edu/naa/features/strong2z.htm)

Discussion

A typology of not just notes, but also disciplinary practice

Field books differ not just between disciplines, but over time; social and scientific practices evolve from one year to the next, and notebooks within in the same domain vary wildly, for example, from 1850 to 1950. While we did find that the significant properties of field notes across these three disciplines fit broadly within the Grace and Knight’s (2008) typology, we also note that the property types have very different implications for the re-use of these records as data sources. For instance, the ‘context’, ‘content’ and ‘behavior’ properties are likely most important for discovery, while the ‘structure’ and ‘rendering’ properties are likely more important for cross-disciplinary collecting and digitization purposes.

Categorizing field notes not just by discipline, but by data structure and data collection method may better support later use of field notes while also informing our understanding of how data collection practices evolve over time. Almost every notebook that we looked at contained data that would be usable by another field; paleontological notes contain descriptions of towns, historical events and people that would be of potential interest to historians and geographers, and archaeological field notes often contain detailed drawings and descriptions of local flora and fauna that would be important to biologists.
Future work

Making use of undiscovered personal knowledge for public science

This work is meant as a first step in exploring how the similarities in research methods and document structures recorded in field notebooks might be formally translated into a better understanding of the significant properties of digitized field notes. In turn, we believe that a better understanding of these properties is important for the success of digitization projects and the development of preservation standards in libraries, archives and museums. We also believe that working towards a taxonomy of field notes and field data will not only aid future efforts in creating a field note markup/annotation schemas, but will also provide important insights to the varied ways that recorded data inform a process of knowledge production.

References


Rafinesque, C. (1818). Notebook kept by Rafinesque on a trip from Philadelphia to Kentucky, 1818. Smithsonian Institute Archives; RU 007250, Box 1 Folder 3.


Identifying Claims in Social Science Literature

Shameem Ahmed, Catherine Blake, Kate Williams, Noah Lenstra, and Qiyuan Liu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign, USA
{ahmed9, clblake, katewill, nlenstr2, qliu14}@illinois.edu

Abstract

The Claim Framework was developed to capture how scientists communicate findings from an empirical study. Although the framework has been evaluated in biomedical literature, the framework has yet to be examined with respect to social science literature. Our goal is to fill this gap and explore the degree to which the Claim Framework can capture claims made in two social science research areas: Community Informatics and Information and Communication Technologies for Development. This poster presents preliminary results on the number and location of claims in full-text social science articles compared to claims in biomedical articles.

Keywords: claim framework, natural language processing, information extraction, community informatics, information and communication technologies for development

Introduction

Social Work Abstract Plus¹ and Sociological Abstracts² are two well-known databases that comprise hundreds of thousands of abstracts from thousands of social science journals. Natural Language Processing (NLP) methods have been explored to identify important concepts contained in technical papers (Paice & Jones, 1993), cause-effect relationships from news group and SIGIR datasets (Mengle & Goharian, 2010) and from newspapers (Khoo, Kornfilt, Oddy, & Hyon Myaeng, 1998) and hypotheses from MEDLINE articles (Srinivasan, 2004). However, little work has explored how well such methods will generalize to a large volume of social science abstracts. Moreover, since abstracts fail to accurately reflect the content of research articles 43 percent of the time (Pitkin, Branagan, & Burmeister, 1999), NLP methods become even more important when we start to consider full-text collections such as ERIC³ and JSTOR⁴.

In 2010, Blake proposed the Claim Framework (Blake, 2010), as a domain-independent representation of how scientists communicate their findings in empirical studies. The framework defines claim as new finding from the articles that brings about an effect or a result. For instance, “Indeed, glycine prevented Wy-14643-stimulated superoxide production by Kupffer cells” is a claim in biomedical literature. On the other hand, “Contrary to much rhetoric, even very poor people chose to have a phone” is a claim collected from social science literature. Although the Claim Framework was developed for the life sciences, such as bioinformatics and clinical informatics, it is not clear how well the framework will generalize to findings reported in the social sciences literature.

Our goal is to explore the extent to which claims made by authors in the social sciences conform to the Claim Framework (Blake, 2010). This poster describes the first step towards that goal by identifying claims in eight full-text articles in two social science research domains: Community Informatics (CI) (Gurstein 2000; Keeble & Loader, 2001; Williams & Durrance, 2009) and Information and Communication Technologies for Development (ICT4D) (Unwin, 2009).
Method

Eight full-text peer-reviewed research articles from two journals of CI and ICT4D (Journal of Community Informatics and Information Technology for Development Journal) were collected. These articles were selected at random from a dataset collected in our earlier study (Williams, Ahmed, Lenstra, & Liu, 2012). Scripts were written to segment the documents into sentences and segmentation errors were corrected. Although different articles contained different number of sections, all eight articles comprised at least five sections, namely, Abstract, Introduction, Research Method, Results, and Conclusion. To standardize, additional sections present in each paper were moved into one of the five categories. Each sentence was annotated either as a claim or as a non-claim.

Results and Discussion

A total of 2,433 sentences remained after sentence errors such as missing spaces after a period, unknown characters, typographical errors, and ambiguous references were corrected. Table 1 summarizes the number of claims made in each article.

<table>
<thead>
<tr>
<th></th>
<th>CI1</th>
<th>CI2</th>
<th>CI3</th>
<th>CI4</th>
<th>ICT4D1</th>
<th>ICT4D2</th>
<th>ICT4D3</th>
<th>ICT4D4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of sentences</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Number of claims</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Number of claims</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Number of claims</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Number of claims</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Number of claims</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Number of claims</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Number of claims</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1
Claims in CI and ICT4D articles

![Figure 1. Percentage of Claims for CI and ICT4D articles](image-url)
Table 2  
Comparative Claims for CI and ICT4D articles

<table>
<thead>
<tr>
<th></th>
<th>CI</th>
<th>ICT4D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average % of claims per article</td>
<td>40.1%</td>
<td>29.8%</td>
</tr>
<tr>
<td>Minimum % of claims per article</td>
<td>35.8%</td>
<td>21.1%</td>
</tr>
<tr>
<td>Maximum % of claims per article</td>
<td>44.7%</td>
<td>38.7%</td>
</tr>
</tbody>
</table>

Figure 1 shows the percentage of claims and Table 2 summarizes the comparative claim statistics for CI and ICT4D articles. The percentage of claims is calculated by dividing the number of sentences that report a claim by the total sentences in an article (excluding references). For example, if there are 100 sentences in an article and among those sentences 34 are considered as claim sentences then percentage of claim for that article is 34%.

Figure 2. Claims for CI articles

Figure 3. Claims for ICT4D articles

Figure 2 and 3 show that the distribution of claims between sections is similar for both CI and ICT4D articles. Here, we see that, authors frequently report claims in the Results and Conclusion sections. However, authors in CI report more results in the Results section than in ICT4D (66.7% vs. 55.1%) and authors of ICT4D articles report more results in the Conclusion section than CI (32.3% vs. 27.0%). The contribution of other three sections, both for ICT4D and CI, are negligible with respect to
claims. These results suggest that Text Mining systems should have more focus on Results and Conclusion section for ICT4D and CI articles, which is quite different than biomedical research domain.

![Figure 4. Claims for Biomedical vs. CI vs. ICT4D articles](image)

As the Claim Framework has already been tested in biomedical articles, we can provide a side-by-side comparison for claims made in biomedical, CI, and ICT4D articles. Figure 4 shows that the proportion of claims made per article is higher in CI and ICT4D articles than in biomedical articles (39.8% vs. 29.5% vs. 22.8%). Authors in biomedical research area tend to have a greater proportion of claim sentences in the abstract section than in social science literature (7.8% vs. 2.1% vs. 5.2%). In addition, authors of biomedical articles have a greater proportion of claim sentences (28.6%) in the Introduction section. In contrast, CI has only 0.9% and ICT4D has 5.5% claims in the Introduction. It seems that while CI and ICT4D authors mainly discuss their motivation, research question, and some background information in the Introduction section, biomedical authors report one-third of their claims in this section.

Another difference was found with respect to the document structure. Only 2 out of the 29 biomedical articles included a conclusion section, whereas all 8 articles in the CI and ICT4D articles included a conclusion section. Sentences in the discussion section of a biomedical article provide the reader with information about context and the implications of the study results, which are important, but differ from the focus of the Claim Framework which includes factual statements about the study findings. In the CI and ICT4D articles, a greater proportion of the sentences in the Result section report claims. As expected, none of the research communities focus on claims in the Methods section (0.5% vs. 2.7% vs. 0.9% for CI, ICT4D, and Biomedical respectively).

**Conclusion**

Several attempts to characterize scientific literature have been made in the life sciences, but little work has been done to explore how well those methods might apply to the social sciences. Our goal is to explore the extent to which Blake’s Claim Framework (Blake, 2010) might apply to social science articles.

Although we have yet to conduct a more detailed analysis of how claims identified in these articles intersect with the five Claim Framework types (explicit, implicit, comparison, correlation, and observation), these preliminary results suggest that there may be differences between social science literature and the biomedical literature with respect to where an author is likely to report study results (claims). Findings reported in this paper have important implications for both information retrieval and natural language processing systems. We plan to increase the number of articles in the analysis as future work and to apply automated methods to identify explicit claims and comparisons described in (Blake, 2010) and (Hoon Park & Blake, 2012) respectively.
References


Participation in the City: A Typology of Open Government Data Use

Morgan Currie
Department of Information Studies, University of California, Los Angeles
msmorgancurrie@ucla.edu

Abstract

Open Government Data (OGD) is defined broadly as any online publication of government documents, and more specifically as government data sets structured for public consumption and reuse. OGD proponents proclaim its benefits: government transparency, increased civic participation, and more efficient, sustainable cities. Critics point out that access to data is not enough: a sophisticated degree of information literacy is required to use OGD effectively for civic agendas. On either side of the debate, what is often missing is a more empirical understanding of who uses OGD and for what purposes. This poster presents preliminary research on methods that might aid in sifting through the civic uses of OGD. I focus on OGD related to city planning in the United States, namely data on air quality, public transit, and housing, to build a classification of OGD usage and to ask how these data sets might be deployed to enhance civic life.

Keywords: participation, open data, literacy, e-government

Poster Abstract

Open Government Data (OGD) is defined broadly as the online publication of any government document or dataset; more specifically it can refer to government data sets that are structured for public consumption and reuse. In the United States the virtues of open government data have historically included greater access to government documents and proceedings, with the aim of airing out government processes by public oversight. This agenda led to the 1966 Freedom of Information Act (FOIA), the first to give citizens the right to request any document from the executive office. Thirty years later the Electronic Freedom of Information Act lets citizens request records in any format made possible by current applications and instructs agencies to publish public records and final policies within a year of creation. Transparency, accountability and access to previously undisclosed information are the classic motivations behind government release of documents such as legislative information, voting records, daily congressional hearing schedules, and state law.

More recently, however, OGD has become an appeal to third-party innovators. Drawing from peer-to-peer practices of source code sharing, many OGD sites put data into machine-readable, adaptable formats protected by open licenses. OGD in this sense promises more than accountability and access. By contributing to a common pool of data for private reuse without restriction, OGD can theoretically allow raw data to accrue unexpected worth. According to the Open Knowledge Foundation, “Part of the beauty of open government data is that it is impossible to predict precisely how it will be used to create value.”¹ Data.gov also spells out that its site goes beyond improving access to Federal data, to increasing “creative use of those data beyond the walls of government by encouraging innovative ideas (e.g., web applications).”²

The implication of OGD’s expanded functionality is that, according, its champions, it will foster new modes of civil participation, bottom-up innovation, and even more efficient public service. Mobile apps, for instance, allow citizens to design a variety of services that can help citizens navigate the city and find resources such as public transit schedules. These apps have made an enthusiast out of Tim O’Reilly, who calls OGD a step beyond e-governance towards ‘Gov 2.0’, or the ‘government as platform’.

¹ From the OKF ‘About’ page at http://opengovernmentdata.org/about/.
² From the data.gov ‘About’ page at http://www.data.gov/about.

Copyright is held by the author.
According to O'Reilly (2009), "In each case, the platform provider raised the bar, and created opportunities for others to exploit." Government information becomes a platform for others to build on, purporting to shift power in the same way that Apple disrupted the phone market when it invited users to design new apps for the iPhone.

There has developed a degree of criticism in response to such optimism for OGD. Professor Michael Gurstein (2011), for instance, points out that participation in OGD more often than not requires technical literacy to repurpose the data usefully and creatively. But how many of us can build an apps and download APIs? The interpretive work essential to using open data effectively – to contextualize it and give it meaning – often requires financial resources, expertise, and training to analyze and report the data to advance certain claims. Gurstein argues that intervention is necessary to make resources for effective use of open data available to low-income and minorities.

Such advocacy would mitigate, to a degree, the problem that in the name of transparency, OGD has become a boon for private commercial investment. Because open licenses don't discriminate against users of raw data, private interests can easily commercialize OGD by appropriating it into non-open systems. The PSI-reuse industry includes LexisNexis, Daily Mail, and Google; Google's Public Data Project, for instance, is currently developing visualization tools that could monopolize data interpretation by locking it into their proprietary, close-source applications. The concern is that without a critical ideological approach to OGD, advocate's zeal for transparency and participation will become another form of corporate subsidy.

On either side of this discussion about OGD's merits and faults, what is missing is more in-depth information gathered on how these data sets are actually being used on the ground to affect change in civic life. Such a study would first need to formulate a typology of usage, whether of a government site's public API for a third-party app, or census data included in a report by a non-profit organization. The research would next ask who is using the data – individuals, civic groups, university researchers or students? – and what this information might say about degree of data literacy required to deploy OGD. Third, it would categorize usage according to a continuum of effectiveness to cause change, so getting to the heart of the question of how much this information encourages civic participation and how affective this participation is. Is the data used to help citizens navigate a transit system through a third-party app, or is it employed to argue for new policies altering a transit system itself? Is the data used to contest policies or representations based on such figures, or even to show where lacunae in public data leads to the suppression of certain demographic information or environmental metrics? Finally, how far can this research go: what methods can catalog the extent of OGD usage in all its forms, and how do we even calculate a metric of OGD 'effectiveness'?

This poster will therefore propose preliminary methods (and obstacles to these) for building a repository of primarily US-based case studies of OGD usage. To narrow its scope (there are after all 389,000 data sets on data.gov alone), I focus particularly on OGD related to city planning in the United States, namely data on air quality, public transit, and housing. The typology will include who is exploiting the data, for what purpose, and how this intervention affects citizen behavior or takes part in political agenda setting. The hope is to present a more complex and informed picture of OGD usage, both for governments and civic organizations or even individuals seeking to understand how they might become more literate in OGD usage, especially as data is increasingly incorporated into the political and economic mechanisms that drive cities today.

References


Lead, Lag or Get Out of the Index: Exploring Macro-economic Indicators of Data Use

Nicholas M. Weber  
nmweber@illinois.edu

Andrea K. Thomer  
thomer2@illinois.edu

Center for Informatics Research in Science and Scholarship  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign

Abstract

Developing techniques to better quantify, track and more accurately describe the impact of federally funded research is quickly becoming a reputable domain for information studies, including data curation. In previous papers we've suggested the adaptation of an existing Data Use Index to quantify data use in the Research Data Archive at the National Center for Atmospheric Research. In this poster we revisit those indicators to determine their ability to forecast or indicate changes in data use over time. Central to our exploration of these indicators is an economic approach to quantifying data use, which holds that patterns in data repository events (downloads, searches, browsing etc.) should be capable of both predicting and explaining variations in use over time. We present preliminary results from this analysis and conclude with some prospects for future work with macroeconomic indicators.

Keywords: data curation, infometrics, scholarly communications, evidence-based policy

Introduction

The transformation of digital datasets from the 'underlying' research material supporting formal journal publications, to first-class, citable and shareable research products has been a gradual process across the social and natural sciences. In recent months this process has sped up considerably, as funding agencies like NSF (Gutman, 2012) have announced that grant applications and policy reports will soon be required cite to research 'products' instead of simply 'journal publications'. This slight shift in language has the potential to profoundly impact the ways we quantify the productivity, impact and usefulness of research funded at a federal level (Lane, 2010; Mayernik, 2012; Parsons, Duerr, Minster, 2010).

Not Metrics of Use, But Indicators of Use

Measuring and tracking the use of datasets still remains largely dependent on the techniques of citation-based bibliometrics, but increasingly there are efforts to diversify both the types of scholarly materials that we measure (Priem and Hemminger, 2010) and the values we assign to such measurements (Piwowar, Carlson and Vision, 2012). In previous work (Weber et al., 2013), we proposed the construction of metrics that 'indicate' data use: Different from traditional measurements of direct citations or acknowledgements, these metrics are constructed from download, browse and search events in a data archive (see table 1). Combined, these indicators make up a Data Use Index (DUI) that allows for a more holistic understanding of how data are used and what impact they have on a community of researchers served by a data archive (Ingwersen and Chavan, 2011). In this analysis we've chosen to explore data use indicators in the Research Data Archive (RDA), at the National Center for Atmospheric Research (for a more thorough discussion of the RDA see Jacobs and Worley, 2009)

Acknowledgements: Thanks to Matthew Mayernik, Steve Worley, Robert Dattore and Zaihua Ji. This research was funded through the DCERC program, IMLS award # RE-02-10-0004-10.


Copyright is held by the authors.
Table 1
Twelve indicators making up a Data Use Index for the RDA.

<table>
<thead>
<tr>
<th>RDA: Indicator of use</th>
<th>Explanation</th>
<th>Type of indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Unique Users (UU)</td>
<td>Unique users that downloaded data per time window</td>
<td>Coincident</td>
</tr>
<tr>
<td>1a Unique Users - Advanced</td>
<td>UUs that accessed data programmatically</td>
<td>Coincident</td>
</tr>
<tr>
<td>1b Unique Users - Assisted</td>
<td>UUs that accessed data via GUI or Service</td>
<td>Coincident</td>
</tr>
<tr>
<td>2 Number of Datasets</td>
<td>Number of Datasets assigned DS number</td>
<td>Coincident</td>
</tr>
<tr>
<td>3 Files DS</td>
<td>Number of files in Dataset per time window</td>
<td>Coincident</td>
</tr>
<tr>
<td>2 Download Frequency</td>
<td>Total number of files downloaded per time window</td>
<td>Leading</td>
</tr>
<tr>
<td>2a Download Frequency - Advanced</td>
<td>Files downloaded by Advanced users</td>
<td>Leading</td>
</tr>
<tr>
<td>2b Download Frequency - Assisted</td>
<td>Files downloaded by Assisted users</td>
<td>Leading</td>
</tr>
<tr>
<td>4 Homepage Hits</td>
<td>Dataset Homepage Hits per time window</td>
<td>Leading</td>
</tr>
<tr>
<td>4a Homepage Hits - Direct Access</td>
<td>Dataset Homepage Hits per time window by users with direct access (link not indexed or retrieved by search)</td>
<td>Leading</td>
</tr>
<tr>
<td>4b Homepage Hits - With Link</td>
<td>Dataset Homepage Hits per time window by users with link (from indexed list or retrieved by search)</td>
<td>Leading</td>
</tr>
<tr>
<td>5 Subset Requests</td>
<td>Subsets Requests per time window</td>
<td>Leading</td>
</tr>
<tr>
<td>6 Download Density</td>
<td>Average number of files downloaded per UU</td>
<td>Lagging</td>
</tr>
<tr>
<td>7 Usage Impact</td>
<td>Total number of downloaded files over total files in dataset</td>
<td>Lagging</td>
</tr>
<tr>
<td>7a Usage Impact - Advanced</td>
<td>&quot;</td>
<td>Lagging</td>
</tr>
<tr>
<td>7b Usage Impact - Assisted</td>
<td>&quot;</td>
<td>Lagging</td>
</tr>
<tr>
<td>8 Interest Impact</td>
<td>Total homepage hits per number of files in dataset</td>
<td>Lagging</td>
</tr>
<tr>
<td>9 Usage Balance</td>
<td>Files downloaded by number of homepage hits per time window</td>
<td>Lagging</td>
</tr>
<tr>
<td>10 Subset Ratio</td>
<td>Number of subset requests over total number files downloaded per time window</td>
<td>Lagging</td>
</tr>
<tr>
<td>12 Secondary Interest Impact</td>
<td>Homepage over UU</td>
<td>Lagging</td>
</tr>
</tbody>
</table>

The Economies of Data Repositories

Vertesi and Dourish first introduced the concept of data economies (2011) to information studies, and we later used this economic lens to explore data use between and within sub-disciplines of researchers involved in Earth Systems Science (Weber et al., 2012). In combining an economic perspective of data work and metric based indicators, we want to ask:

- Can we use macro-economic indicators to predict future data use in an archive setting?
- Can we explain patterns in data use which vary from temporal norms?
- Are there data use events that we’ve not collected that could better predict use and access patterns over time?
Macroeconomic Indicators

In traditional macro-economic analysis, there are many types of indicators (e.g. Performance, Technical, Directional or Temporal etc.), but most common for investigating broad, temporal trends are the rather recursively named ‘Economic Indicators’ (Moore, 1983). For this analysis, we’ve chosen to analyze search and download events from datasets hosted by the RDA using the three most common temporal Economic Indicators:

- **Leading indicators** are used to forecast how slight changes in patterns of use or disruptions / enhancement in access to data may in fact foreshadow larger shifts in the ‘economy’ of the data archive.
- **Lagging indicators** are usually post-hoc calculations that attempt to correlate metrics with how or why a certain event happened the way that it did. When used comparatively with leading indicators, lagging indicators can substantiate or refute predicted changes in an economy.
- **Coincident indicators** give a snapshot of the here and now of an economy. Traditionally coincident indicators are representative of a current economic state not yet affected by leading indicators, and not yet represented by lagging indicators. In this sense, the leading and lagging indicators give a sense of certainty to coincident indicators.

Analysis of RDA Indicators

Method

We first categorized our 12 usage indicators according to their corresponding economic indicators (see table 1, column 4). We expect that discovery- and access-related events will act as leading indicators; simple data such as the number of registered users for the archive, or the amount of files within a given dataset will act as coincident indicators; and lagging indicators will combine these leading and coincident measures to explain patterns of use, and measure the relative impact of a dataset from the RDA.

We then calculated these 12 usage indicators for three of the most heavily used datasets hosted by the RDA; these also represent the diversity of the RDA’s holdings, as they include various types of climate data, such as model output data, reanalysis data, and observational data collected by field campaigns. It’s important to note that while each of these datasets are assigned one unique identifying number (e.g. ds083.2), they are composed of numerous files, which users often subset or download in various combinations. After calculating the indicators listed in table 1, we then graphed leading, lagging and co-incident indicators for each dataset to explore shifts or variations in data use in the RDA.

Preliminary Results

Some of our most interesting results have been within our leading indicators: Increases in download frequency by type of user access often has relationship with total amount of data downloaded per month. So, months in which programmatic (advanced) user access rises we see a much higher overall download frequency- whereas increases or decreases in assisted user access seems to negligibly affect the total download frequency.

This seemed puzzling given that we expected assisted user access, which is aided greatly by the data curators at NCAR, to positively effect the overall amount of data consumed by archive users. However, our leading and co-incident indicators revealed a an interesting relationship between the number of subset requests and advanced user download rates. As sub-set requests increased the overall download frequency decreased. So, the implication might be that user services (like sub-setting) actually lead to more selective data use, and overall that assisted access might actually decrease the total amount of data downloaded in a given month. One conclusion we might draw from this analysis is that data archives promoting their success using a raw metric such as the total amount of data downloaded per month may actually be confusing the inefficiency of their architecture, with their impact on a user community.
Future Work

Our analysis is ongoing, but future work will include a more thorough analysis of the data available from the RDA, as well as an exploration of the predictability of these indicators in the form of lagging indicators. This work is also a first attempt at understanding use and impact amongst federally funded research data. We expect that as more datasets and archives are analyzed, these indicators will become more stable, and better able to accommodate an economic modeling of the impact of research data.

References


A Model for Assessing the Quality of Gene Ontology

Shuheng Wu
Florida State University
College of Communication and Information
sw09f@my.fsu.edu

Abstract

With the proliferation of bio-ontologies in the molecular biological community, concern remains about their quality. There are a number of frameworks and models have been developed to assess the quality of Gene Ontology, which is one of the most influential and widely used bio-ontologies in the community. However, without any theoretical guide for quality assessment, most of these frameworks and models are incomplete and intuitive, and cannot assess the quality of Gene Ontology consistently, systematically, and completely. This study uses a theoretical information quality assessment framework to guide the development of a quality evaluation model for the Gene Ontology, using both the conceptual and empirical approaches.

Keywords: data quality, quality evaluation, knowledge organization, ontologies, Gene Ontology

Introduction

Since the publication of human genome in 2001, the world has entered into the “post-genome age” (Higgs & Attwood, 2005, p. 4). The vast growth in the amount of biological sequence data has led to the coexistence of heterogeneous data types, formats, and vocabularies delaying research process in biology and posing challenges for data management in biological repositories. There are urgent needs in scientific communities for knowledge organization (KO) systems (e.g., metadata schema, ontologies, Semantic Web) to provide access to and make sense of huge amount of scientific data (Gray, 2007; Gray et al., 2005). Scientific research has become increasingly multi-institutional, multinational, and interdisciplinary. This creates challenges for traditional KO systems to represent interdisciplinary data and improve data and metadata interoperability across disparate vocabularies and domains (Allard, 2012). It also raises the expectation on the quality of scientific data and its metadata (Anderson, 2004).

Due to the complexity of molecular biological entities (e.g., genes, proteins) and their relationships, there has been a trend towards the development and adoption of bio-ontologies in the biomedical and molecular biological communities. Among many current bio-ontologies, the Gene Ontology (GO) is one of the most influential in molecular biology and biomedicine, and has been widely used for text mining and information extraction (Blaschke, Hirschman, & Valencia, 2002). The difficulties of maintaining ontologies are in gaining community acceptance, integrating new knowledge, and reflecting established knowledge (Open Biological and Biomedical Ontologies, 2011). With the proliferation of bio-ontologies, concern remains about their quality.

Köhler, Munn, Rüegg, Skusa, and Smith (2006) proposed two automatic metrics—circularity and intelligibility—to assess the quality of term definitions in ontologies, and tested the metrics using empirical data collected from GO. Buza, McCarthy, Wang, Bridges, and Burgess (2008) developed a composite automatic quality metric—GO Annotation Quality (GAQ) score—to evaluate the quality of GO annotations, and tested it by measuring the annotations for chicken and mouse over a period of time in GO. The GAQ score is a product of the breath of annotation (i.e., the number of GO terms assigned to each gene product) and the evidence code (i.e., an indicator of the source of annotation) rank.

Acknowledgement: Many thanks to Dr. Besiki Stvilia for helpful conversations.
Copyright is held by the author.
Leonelli et al. (2011) collected empirical data from GO curators, and identified five quality problem sources: (a) mismatch between GO representation and reality; (b) the scope extension of GO; (c) divergence in how the GO terminology is used across user communities; (d) new discoveries that change the meaning of GO terminology and relationships; and (e) the addition of new relations. Defoin-Platel et al. (2011) proposed a framework of 12 quality metrics for assessing the quality of GO’s functional annotations. Most of these frameworks and models are incomplete and intuitive, mainly based on individual perception of quality requirements. Without any theoretical guide for quality assessment, they are unable to assess the quality of GO consistently, systematically, and completely. With the increasing popularity and impact of GO in the molecular biological community, there is a great need for developing a comprehensive and systematic quality evaluation model for GO to inform KO system designers and curators and support users’ decision making process.

### Theoretical Framework

This study uses Stvilia’s theoretical Information Quality (IQ) assessment framework (Stvilia, Gasser, Twidale, & Smith, 2007) to develop a quality evaluation model for GO. Stvilia’ IQ assessment framework consists of a well-defined typology of IQ problem sources linked with affected information activities and a taxonomy of 22 IQ dimensions along with 41 generic IQ metrics. The reason for using Stvilia’s framework is it provides consistent and complete logic to deal with context sensitivity, specifying methodologies to analyze an information activity system and identify users’ IQ value structure and allowing rapid and inexpensive development of context-specific IQ measurement models. The generic IQ metrics have been successfully reused in different contexts (e.g., the English Wikipedia). Compared to previous IQ assessment frameworks, Stvilia’s is more systematic, comprehensive, and reusable (Stvilia, 2006; Stvilia et al., 2007). Stvilia’s framework has been operationalized in different settings (e.g., an online collaborative encyclopedia, an aggregated digital repository) and domains (e.g., biology, healthcare, and information science). Stvilia (2007) used his framework to construct a model for assessing the quality of biodiversity ontologies, and suggested future research to develop quality evaluation models for type-specific ontologies.

### Methodology

Following Stvilia and Gasser’s (2008) suggestion, this study uses both conceptual (top-down) and empirical (bottom-up) approaches, since a conceptual IQ model can guide the empirical analysis but may differ from a community’s active IQ model and empirical data can reflect the community’s actual IQ requirements. The conceptual approach is built upon an information entity’s use scenarios to analyze its activity system context (Stvilia et al., 2007). This study uses activity theory (Leont’ev, 1978; Nardi, 1996; Vygotskii, 1978) and the findings of a related study (Wu, Stvilia, & Lee, 2012) to guide the conceptual analysis to identify the activities of using, developing, and maintaining GO, and the types of quality problems to which GO may be prone. The conceptualization of GO’s activity system context and the suggested quality problem structure will be used to guide the empirical analysis to develop a quality evaluation model for GO.

The empirical approach involves qualitative and quantitative analyses on the Ontology and GO users’ IQ evaluations, generating statistical profiles of GO and users’ IQ value structure (Stvilia et al., 2007). GO has created the curator requests tracker to allow users to provide feedback to the Ontology, such as suggesting a new term or definition and reorganizing a section of the Ontology (Gene Ontology, 2012a; SourceForge, 2012). GO curators review users’ requests, and implement edits where appropriate. GO also has a mailing list—GO Discuss—for users to report errors or omissions in GO annotations (Gene Ontology, 2012a, 2012c). Next in the empirical analysis, this study will collect GO users’ requests and curators’ comments from these sources to conduct content analyze to identify GO community’s quality requirements. Finally, this study will relate the identified quality problems to the activities around and the measurable attributes of GO and GO annotations to develop a set of context-specific IQ metrics to evaluate the quality of GO.
Preliminary Findings

Go consists of three ontologies describing the cellular locations, molecular functions, and biological processes of genes and gene products in a species-neutral manner, and intends to provide each gene and gene product with a cellular context (Gene Ontology, 2012a). The development and maintenance of GO consist of three parts: (a) developing and maintaining GO terms and relationships among the terms; (b) annotating gene products, associating genes and gene products curating in collaborating databases (e.g., UniProt, WormBase) with GO terms; and (c) developing tools to facilitate the development, maintenance, and use of GO. Users can access GO terms and annotation data via a browser named AmiGO.

There are generally two types of annotation in GO: computational and manual (Gene Ontology, 2012b; Rogers & Ben-Hur, 2009). Computational annotation usually involves: (a) searching for a similar gene to a newly sequenced gene in BLAST; (b) searching AmiGo to find GO terms associated with the similar gene; and (c) assigning those GO terms to the new gene, assuming similar genes share similar cellular context. Biologists usually perform manual annotation through laboratory experiments to learn about the cellular context of genes and gene products. Manual annotation is accurate but time-consuming and labor-intensive. Compared to manual annotation, computational annotation is fast but less accurate and detailed. Each GO annotation contains an evidence code, indicating whether the annotation is inferred from primary research, literature, curators, computational analysis, or other databases.

Using three data use scenarios, a recent study identified and conceptualized the data quality problem sources in molecular biology as incomplete and inconsistent mapping, and dynamic quality problems caused by context changes, changes in the entity, and changes to the underlying entity (Wu et al., 2012). The data quality change model developed in that study may be applicable to GO as the concepts represented by GO are within the domain of molecular biology.

Based on the GO annotation activities, one may expect GO annotations have inaccurate, incomplete, and inconsistent mapping problems. An inaccurate/incomplete computational annotation is an instance of inaccurate/incomplete mapping between a gene product and GO terms. Inaccurate definition of a GO term may also cause inaccurate mapping between the GO term and gene products. An inconsistent mapping may occur when a GO collaborating database updates the cellular context of a gene product (e.g., new discovery of protein functions), but its GO annotation remains unchanged. One may also expect dynamic quality problems in GO terms and annotation data. Over a time period, there might be new discoveries that change the meaning of certain GO terms, and might change their relations to other GO terms and the annotation data. Similarly, there might be new discoveries of the attributes of gene products, which might change their GO annotations.

Conclusion

Guided by Stvilia’s theoretical IQ assessment framework, this study aims to construct a model to assess the quality of GO using both conceptual and empirical approaches. The conceptual analysis of this study identifies the activities of creating, using, and maintaining GO, and suggests types of quality problems may occur in GO. This conceptualization will inform the empirical analysis, the next step of this study, to develop GO community’s active IQ evaluation model.

References


Improving the Character of Optical Character Recognition (OCR): iDigBio Augmenting OCR Working Group Seeks Collaborators and Strategies to Improve OCR Output and Parsing of OCR Output for Faster, More Efficient, Cheaper Natural History Collections Specimen Label Digitization

Robert Anglin  
North American Bryophyte and Lichen TCN/Symbiota  
ranglin@wisc.edu

Jason Best  
Botanical Research Institute of Texas (BRIT)/Biodiversity Informatics  
jbest@brit.org

Renato Figueiredo  
University of Florida/iDigBio  
renatof@ufl.edu

Edward Gilbert  
North American Bryophyte and Lichen TCN/Symbiota  
egbiodiversity@gmail.com

Nathan Gnanasambandam  
Xerox Research Center Webster  
nathan.gnanasambandam@xerox.com

Stephen Gottschalk  
New York Botanical Garden  
sgottschalk@nybg.org

Elspeth Haston  
Royal Botanic Garden Edinburgh  
e.haston@rbge.ac.uk

P. Bryan Heidorn  
University of Arizona/School of Information Resources and Library Science  
heidorn@email.arizona.edu

Daryl Lafferty  
Arizona State University/SALIX  
daryl@daryllafferty.com

Peter Lang  
ABBYY USA  
peter.lang@abbyyusa.com

Gil Nelson  
Florida State University/Institute for Digital Information (iDigInfo)  
gnelson@bio.fsu.edu

Deborah Paul  
Florida State University/Institute for Digital Information (iDigInfo)  
dpaul@fsu.edu

William Ulate  
Missouri Botanical Garden/ Biodiversity Heritage Library  
william.ulate@mobot.org

Kimberly Watson  
New York Botanical Garden  
kwatson@nybg.org

Qianjin Zhang  
University of Arizona/School of Information Resources and Library Science  
gianjinzhang@email.arizona.edu

Abstract

There are an estimated 2 – 3 billion museum specimens world – wide (OECD 1999, Aríño 2010). In an effort to increase the research value of their collections, institutions across the U. S. have been seeking new ways to cost effectively transcribe the label information associated with these specimen collections. Current digitization methods are still relatively slow, labor-intensive, and therefore expensive. New methods, such as optical character recognition (OCR), natural language processing, and human-in-the-loop assisted parsing are being explored to reduce these costs. The National Science Foundation (NSF), through the Advancing Digitization of Biodiversity Collections (ADBC) program, funded Integrated Digitized Biocollections (iDigBio) in 2011 to create a Home Uniting Biodiversity Collections (HUB) cyberinfrastructure to aggregate and collectively integrate specimen data and find ways to digitize specimen data faithfully and faster and disseminate the knowledge of how to achieve this. The iDigBio Augmenting OCR Working Group is part of this national effort.

Keywords: iDigBio, OCR, natural language, information analysis, machine language


Copyright is held by the authors.
Introduction

While optical character recognition (OCR) is currently utilized by some museums in their databasing workflows, better OCR strategies would increase the chances of meeting the following goals. Part of iDigBio’s mission is to assist the biodiversity collections community in finding ways to:

- speed up the overall digitization process,
- lower the cost,
- improve overall efficiency,
- assure digitized data is fit-for-use (NIBA 2010, Chapman 2005), and
- provide the resulting digitized data records to researchers more quickly.

Some Projects and Challenges of the A-OCR Working Group

Those currently using OCR note there is also much room for improvement in issues including parsing of the output, autocorrection of text, recognition of text, recognition of handwriting and image segmentation. The iDigBio Augmenting OCR (A-OCR) working group, formed in March of 2012, is actively engaged in identifying opportunities to leverage OCR tools and technologies that are successful (both within and outside of the biology digitization domain) and disseminate these tools, methods and workflows to the public. The A-OCR working group would like to integrate these tools, or seek funding for tool development.

Natural history museums contain a wealth of specimen data currently only accessible to those with the time, resources and permissions necessary to travel to the museums and walk through the research collections. Since most inventories are not accessible via the web, it is difficult for a researcher to ascertain where important specimens might exist. Collections vary in size from a few thousand specimens in research universities to many millions in the major natural history museums of the world. As part of a national, coordinated, multi-faceted effort to collate, integrate and expose this so-called "dark data" through a cyberinfrastructure hub, the National Science Foundation (NSF) started the Advancing Digitization of Biodiversity Collections (ADBC) program which then funded Integrated Digitized Biocollections, or iDigBio, to build this cloud-based database resource.

The data comes from NSF-funded Thematic Collection Networks (TCNs). The TCNs, made up of groups of museums, are funded to collect data from defined specimen groups in order to address specifically-proposed, timely research themes such as global warming and climate change, species discovery, and species-host-parasite relationships. Besides building an agile cloud-based system to facilitate synthesizing diverse museum collection data sets for research, iDigBio’s goals include working with TCNs, natural history collections, and the broader community to look for ways to produce fit-for-research-use research data quicker and cheaper.

Since much of the to-be-captured data resides on museum specimen labels or in field notebooks as print, type-written text or hand-writing, OCR, algorithms for parsing OCR output, and efficient user interfaces for these tasks are natural targets for improvement in attempts to hasten data capture and insertion of that data into databases. The iDigBio Augmenting OCR Working Group (A-OCR) formed in March of 2012 and after outlining possible goals, held its first workshop on October 1 - 2, 2012 in Gainesville, Florida to:

- build a strategic plan for broader community engagement in our endeavors,
- combine our collective knowledge and experience with current OCR software and parsing strategies to produce website content at iDigBio for use by anyone seeking effective OCR practices when digitizing museum specimens,
- choose hackathon goals for our first iDigBio Augmenting OCR hackathon being held and hosted at the Botanical Research Institute of Texas (BRIT) concurrent with this 2013 iConference, and
- learn about recent developments in OCR, handwriting recognition, and OCR output parsing from the broader community and our working group members.

Each member of our working group brings knowledge and experience from unique uses of OCR and OCR output. As a group, we collected all the issues we would like to work on, for example: improving automated image segmentation. This involves identifying the text block in complex images such as an herbarium specimen or a full tray image of insects. The sample herbarium sheet image in figure 1 (Figure 1) exemplifies the complexities of the task. Here the goal would be to develop an algorithm that quickly
and correctly recognizes the label and ignores the plant. This would enable OCR of these objects to skip image-processing steps currently used like taking a separate image of just the label or using humans to crop the image by hand or indicate (segment) where the label is on a sheet.

Figure 1. Herbarium Sheet, Florida State University, Robert K. Godfrey Herbarium. Used with permission.

Another issue of interest involves developing algorithms that differentiate and classify image segments by successfully figuring out which section contains the primary label, the annotation label (if any), the herbarium stamp, the collecting event label (refers to insect specimens), or other text that may exist on the specimen. Once recognized, segmented OCR output is parsed into fields based on a data standard like Darwin Core for automated insertion into a database.

Only some label types, mainly those printed, and some typed, result in OCR output suitable for this type of parsing. Here’s an example of such a label (Figure 2) and its parsed data.

Figure 2. Label suitable for effective OCR. Herbarium of Yale University. Used with permission.
The North American Bryophyte and Lichen TCN (LBCC) has a goal of digitizing 2.3 million lichen and bryophyte specimens representing well over 90% of North American specimens. To achieve this goal, LBCC has integrated OCR and NLP capabilities into their processing workflows and their Symbiota web portals. Symbiota (http://symbiota.org) is open source software designed to aid biologists in establishing specimen-based public data portals. LBCC is making use of a suite of specimen management tools integrated into the basic user interface (Figure 3) that supports the digitization of specimen information directly from the images of the specimen labels (Figure 4).

Figure 3. Symbiota user interface. Note display of data record, image of label and ocr output.
While OCR, NLP, duplicate harvesting, and concepts of crowdsourcing have been integrated into the working model, the LBCC project continues to work on increasing efficiency and improving performance of these tools.

The Apiary Project (http://www.apiaryproject.org/) is a collaborative effort between the Botanical Research Institute of Texas and the Texas Center for Digital Knowledge (http://txcdk.unt.edu/) at the University of North Texas with the goal of providing a high-throughput workflow for computer-assisted human parsing of biological specimen label data. The Apiary workflow utilizes a three-stage process for extracting parsed text from digital images of herbarium specimens. This workflow provides a user interface through a web-based application. In the first stage, users view the full specimen image and delineate and classify image regions that contain textual content (Figure 5).
In the next phase, these regions are processed by three OCR processes and the user is able to select the most accurate output. When the text output is not accurate, the user may make corrections or, as often is the case with handwritten labels, disregard the OCR output and transcribe the complete text of the region (Figure 6). Once the transcription is complete, the text is parsed into Darwin Core fields (Wieczorek et al., 2012) using controlled vocabularies and interface devices to help standardize and normalize the parsed record.
Next, a key aspect of the iDigBio cyberinfrastructure is the ability to provide cloud-oriented services to its users. In the context of OCR workflows, these services can include common Web-based services hosted by iDigBio and academic or commercial partners, as well as providing users and developers with the ability to develop, configure, package and disseminate new and experimental services by creating virtual appliances. Virtual appliances are pre-configured, ready-to-use “virtual machines” that include all the complex software and configuration needed for an OCR tool or workflow (operating systems, applications, libraries, scripts, etc) in a manner that allows the appliance to be instantiated by end users on their own computers, and/or hosted in the iDigBio cloud infrastructure.

**Conclusion**

We actively encourage you to contact any member of the iDigBio Augmenting OCR working group to get involved. We need your collective energy and knowledge, from graduate students, programmers and professors to commercial companies ~ all are needed and welcome. Comments and collaboration anticipated and appreciated!

Acknowledgements: iDigBio is graciously funded by a grant from the National Science Foundation’s Advancing Digitization of Biological Collections Program (#EF1115210). To each and every member of the iDigBio Augmenting Optical Character Recognition working group, many kind thanks. All of us in the working group would like to express our gratitude to the 2013 iConference for the education and outreach opportunities given to our working group at this gathering of the iSchools community.
References


Using Etherpads as Platforms for Collaborative Learning in a Distance Education LIS Course

Bob Pymm  
Charles Sturt University, Australia  
rpymm@csu.edu.au  

Lyn Hay  
Charles Sturt University, Australia  
lhay@csu.edu.au

Abstract

This poster presents findings of a project which explored the online collaborative learning experiences of distance education (DE) students undertaking an undergraduate information studies. The School of Information Studies (SIS) at Charles Sturt University (Australia) is a leader in the provision of library and information studies education by distance. A recent review of undergraduate and postgraduate courses resulted in greater integration of Web 2.0 technologies to support DE teaching and learning. This included the redesign of the collection management course which serves approximately 100 undergraduate students per year. The aim of this project was to encourage greater communication and collaboration between students to address the challenges of isolation and lack of student community reported by previous cohorts. The researchers found the use of a collaborative document sharing platform, Etherpad, an effective tool in developing students’ skills in critical thinking and collaborative practice, leading to an enhanced DE learning experience.

Keywords: distance education, online collaboration, group-based assessment, etherpad, collaborative document sharing

Introduction

Within Australia, one of the largest providers of distance education is Charles Sturt University (CSU). With 20,000 students enrolled in DE courses, it has a commitment to excellence in the development and delivery of online education (Charles Sturt University, 2012a). The university, like many others, also aspires to produce well-rounded graduates who have the capacity to not only contribute to their chosen professional field but also more broadly to the wider community through the implementation of CSU’s ‘Statement of Graduate Attributes for Undergraduate Courses’ (Charles Sturt University, 2012b).

One of the leaders in online distance education (DE) at CSU is the School of Information Studies (SIS), which has offered DE courses for nearly 30 years. A review of SIS programs was undertaken during 2008/9, and as a result new courses (all offered only in DE mode) were introduced to take advantage of the affordances offered by the online, particularly Web 2.0 environment (Hider et al, 2011). Distance education, by definition, creates a number of challenges for lecturer and student in building and maintaining connection and commitment. The challenges that need to be overcome include communication difficulties, lack of student motivation, high drop out from courses, provision of support at a distance, and a sense of isolation or lack of student community (Crease, Pymm & Hay, 2011). Moody (2004) highlights issues such as the mode of delivery, difficulties in establishing a learning community, and a loss of personal contact which combine to diminish the value of distance learning and contribute to high attrition rates.

Bearing these challenges in mind, a number of strategies have been developed and implemented since 2010 when the revised courses were first made available, to help ‘connect’ students more fully to their cohort and their studies more broadly. One of these strategies, the use of the collaborative document sharing software Etherpad, was introduced into a large undergraduate course with the broad aim of encouraging greater communication and collaboration between these geographically dispersed students. In addition, the use of such software facilitated the introduction of a group-based assignment, traditionally more challenging to implement for distance students but bringing benefits in terms of collaborative learning (Timberlake, 2010).

Copyright is held by the authors.
Etherpad enables participants to simultaneously update a document, providing a tracking mechanism in the form of a “time-slider” to enable the reader to move back and forth over the life of the document—a particularly useful tool for faculty in assessing contributions to the evolution of the finished document. It also enables each author’s contribution to be shown in a different colour, thus again, providing a ready way for faculty to see who has contributed what to the document. Adjacent to the text document, Etherpad also provides a Chat function sidebar which captures a record of conversation between group members while collaborating on document content.

Etherpad is open-source with the code freely available for download and installation (Etherpad Foundation, nd). This was the approach taken at CSU where the software has been installed on a local server, requiring minimal support in-house.

The Research Question

The aim in making use of the Etherpad software was broadly two-fold: to encourage greater communication and collaboration between distance students in order to help develop their capacity to meet a number of the University’s graduate attribute outcomes; and to provide the opportunity to undertake an authentic collaborative learning task that would support discipline specific learning as well as develop more generic skills. The research therefore aimed to assess how well these goals had been achieved.

Methodology

For two years, 2010 and 2011, participants in an undergraduate collection development class (each year comprising around 100 students) were given a collaborative assignment to undertake using Etherpad. Each year students were randomly allocated to groups of four, given the link to a blank Etherpad document, and asked to work together in examining a particular collection development policy, commenting on its strengths and weaknesses. They were also encouraged to break down the policy and negotiate the allocation of workload within their group so that each person was responsible for a specific aspect. All discussion regarding the planning and management phase of the group task was to be undertaken in the Chat sidebar of Etherpad, although the final (assessable) evaluation was to be created as the Etherpad document itself. Students were assessed individually for their overall contribution to the final response, with a mark allocation given for leadership and organization as indicated through the Chat discussions.

The transcripts of Etherpad Chat sessions of 113 student groups were collected across four teaching sessions in 2010-2011. A selection of group conversations were analysed for discrete themes and attributes, with the following five themes emerging as the predominant functionality of the Etherpads as an online collaborative platform:

- The use of Etherpad as a knowledge and values building platform to support the development of discipline-based knowledge (library and information science concepts, principles and practices) and knowledge and values of global citizenship (national and international perspectives, inclusivity, sustainability);
- its use as a social platform (getting to know each other; sharing details of their lives);
- its use to provide affective support (helping with anxiety, stress, fear, uncertainty, reassurance);
- its use as a problem solving platform (students helping each other understand concepts and issues, solve problems, working together to critically evaluate and analyse policy content and issues); and
- its use as a project management platform (dealing with the logistics of the group-based assignment; negotiating task allocation; organizing meeting times).

The university-wide Graduate Attributes document was also used to identify additional attributes which resulted in the development of a coding schedule comprising 15 specific behaviours. Table 1 overleaf presents the scope of each behavior code and the grouping of these codes according to the above five (5) functions as platforms.

All 113 transcripts were then analysed according to these codes.
Results

Analysis of the sessions showed there was a high level of social interaction engendered through the Chat function with most students quickly building a rapport within their group as a result of being task-focused. The majority of groups used the requirements and demands of the shared task to readily establish a community of practice approach that supported collaborative decision making. Only a handful of groups had difficulty with this initial community building phase of the group process. This was either due to individuals within a group misunderstanding what was required to effectively build a sense of collaboration and community within their group, or remaining members of a group who had to be merged with other individuals whose group size had diminished due to course attrition early in the session.

The number of postings per team to the Chat sidebar varied greatly, from 0 to 584, with a median of around 68 postings. With the majority of postings being of a social nature, this generated a supportive environment which clearly assisted in problem solving and helping group members ‘stay on task’. It greatly increased the personal contact between individual students within the class, in comparison to courses not utilising group-based tasks or synchronous communication. Overall, the groups’ conversation captured in the chat transcripts suggests an improved sense of connection with others, and with the course more generally. Whether this will transfer to greater identification and commitment to the program is yet to be determined.

Results related to each theme are summarised below.

As a knowledge and values building platform. Students were assigned randomly to groups so that it was quite likely that within a group there could be a member with considerable experience or knowledge, or someone with absolutely no knowledge other than that gained from the course. Thus building specific discipline knowledge through sharing was a widespread occurrence. Examples of comments included:

“what performance measures/statements have been included in the policy to allow for public accountability. e.g. How can the public measure if the policy is being met.”

“Well if you need help with anything, just ask via email. I've been in academic libraries for a few years now.”

“I guess we didn't discuss the importance of the conspectus and perhaps made assumptions. Big learning curve for working in groups!”

Evidence of thinking more broadly and linking to wider graduate values and attributes was less commonly expressed. This could be due to limited or less explicit presentation of values-based content or lack of learning tasks dealing with values within the course. That said, comments below suggest that there was some consideration of these aspects.

“No no, it's fine we'll all need to learn about it because it will be/effect every aspect of the policy.”

“Hi I work in a public library in Sydney. The only difference is that this library has a big collection of community language.”

“The policy mentioned 3% population is indigenous Australian but the policy does not mention service to meet this community's needs.”

As a social platform. Overwhelmingly, this collaborative work served to provide a foundation for extensive social interaction. From the initial introductions to each other, through discussions on the novelty of this learning approach, to consideration of the assessment requirements, over 50% of groups exhibited a high level (40+ messages) of friendly, social interaction. Thus comments representing building of rapport and familiarity between group members such as these were common:

“Time for my bedtime..getting old.lol. Nice chatting with you. Catch up with you later.”

“yeah I was just reading. I'm glad you guys haven't been doing much either”

“Everyone knows everyone!!! which is good in some ways! Next time you see xxx please tell her I said Hi and hope she is feeling better.”

The comments varied in range and intensity across groups but for the high level ‘interacters’, this positive, friendly tone was engendered at the beginning and continued throughout the assessment work.

As a platform for affective support. This was strongly evidenced across the majority of groups illustrating the desire of DE students to both give and receive affective support as part of the study experience:
“Yes hopefully I will get faster at my work, I lack confidence, its all a bit overwhelming! its been a long time since I studied.”
“I've just read both of you paragraphs so far, wow! I'm impressed!”
“I meant it when I said i thought you did a good job on your article :-(”
“I'm all done people. I had fun with this. Good luck to you all :)”
“no, of course you are not here, its four days later! Muchos apologies colleagues, I will log on tomorrow evening.”

As a problem-solving platform. Numerous enquiries were made regarding use of the software, interpretation of the assessment task and generally seeking reassurance that everything was on the right track. Thus comments such as those below were common place.

“Q: do we have to evaluate in to 300 words; A: think it is 250 approx. You have 500. :)
“no submission. It just closes 22nd and then bob follows link checks it out. you cant submit it via Easts anyway”
“I have downloaded it but still figuring out how to use it!! Can you put the actual references in or just the citation?”

Furthermore, discussion between students in a number of the groups illustrated the contribution of the individual to a group’s development of critical and reflective judgement.

As a project management platform. While task-oriented groups used the Etherpad efficiently as a project management platform, a high level of social interaction in some groups hindered more direct, organizational type discussions. Overall, discussions on the selection of a policy to study, methods of presenting the discussion, and responding to the lecturer’s directions, tended to be highly democratic, seeking input from all. While it was unusual for one person to take a directive tone, it was common for one or two students to emerge as ‘leaders’ in the early stages of community building with the group, with leadership styles being inclusive as opposed to a dictatorial approach. Across all 100 or so active groups, none included any strong directive discussions. Typically comments were couched to be helpful in progressing the project but to avoid being ‘pushy’. Thus:

“Yes, xxx, I agree with you. I think you can work on collection management, weeding, acquisition, etc. Hope you can make up your mind.”
“That's us settled then! A: Government publications, P: Digital resources, J: Retention & withdrawals, and B: Popular culture. Is that right? If so, let's get started!!!”
“are we all happy to leave it as is? or does anyone want to organise it? i don't mind either way”

Conclusion

A major aim of developing this collaborative assignment task was to encourage engagement, communication and critical thinking amongst students studying at a distance who traditionally have found group work challenging; and to assist in making those vital connections to help form a sense of engagement with their peers and their course. The researchers concluded that the nature of the task, together with the technology employed, made a considerable positive impact on those involved, increasing their sense of being part of a cohort, encouraging a questioning, supportive environment and making them feel more at ease with group work. Further analysis will be undertaken into the wealth of data collected in the Chat Rooms to help reveal the extent of the connections made and ongoing study should reveal the degree to which this has translated into improved student retention and identification with the program.

This experience has demonstrated the power of such an approach to community building for DE students as confirmed by this student:

“Hi! A, B and C. Just finished my etherpad( part1 & part 2).it was great working with u all.all the best with rest of the course.”
Table 1: Coding schedule according to types of platforms and behaviours

<table>
<thead>
<tr>
<th>Knowledge &amp; value building platform</th>
<th>Social platform</th>
<th>Affective support platform</th>
<th>Problem solving platform</th>
<th>Project management platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIS</td>
<td>CH</td>
<td>AFF</td>
<td>AS</td>
<td>PM</td>
</tr>
<tr>
<td>LIS demonstate a broad overview of their field, i.e. LIS discipline knowledge</td>
<td>For incidental chat, e.g. hello, bye, etc</td>
<td>use of etherpad chat to provide affective support, e.g. fear, concern, anxiety, happiness, satisfaction, etc</td>
<td>demonstrate analytical skills, including the exercise of critical and reflective judgment</td>
<td>project management tasks &amp; discussion on planning tasks, e.g. dealing with the logistics of the group-based assessment task, who has done what, who hasn’t done what, task allocation, etc</td>
</tr>
<tr>
<td>COM</td>
<td>SO C</td>
<td>PS</td>
<td>PS</td>
<td>ASS</td>
</tr>
<tr>
<td>COM communicate effectively using the language of the discipline, e.g. discussion of collection development concepts, issues etc</td>
<td>Social conversation incl. getting to know each other as people outside university study, sharing details of their life/work/family/partying</td>
<td>address unfamiliar problems; conversations where students help each other solve problems, gain an understanding of a concept or issue</td>
<td>Assessment related discussion in terms of task requirements, management &amp; completion</td>
<td></td>
</tr>
<tr>
<td>VALUES</td>
<td></td>
<td>IT</td>
<td>PL</td>
<td>LEA D</td>
</tr>
<tr>
<td>VALUES demonstrate an understanding of, &amp; commitment to, values-driven practice in their field of study</td>
<td></td>
<td>Technical difficulties related to Etherpad &amp; other software</td>
<td>Peer learning, e.g. acknowledgemen t that they can learn off each other</td>
<td></td>
</tr>
<tr>
<td>INT</td>
<td></td>
<td>ETH</td>
<td>LEA D</td>
<td>Examples of student leadership of group</td>
</tr>
<tr>
<td>INT demonstrate a national and/or international perspective</td>
<td></td>
<td>Discussion on use Etherpad incl. its features &amp; functionality</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Poster presentation

A copy of the poster presentation is available at
http://studentslearn.wordpress.com/presentations/iconf2013/

References


SoMe Tools for Social Media Research

Shawn Walker
stw3@uw.edu

Jeff Hemsley
jhemsley@uw.edu

Josef Eckert
jeckert@uw.edu

Robert Mason
rmmason@uw.edu

Karine Nahon
karineb@uw.edu

University of Washington

Abstract

The Social Media Lab at UW (SoMeLab.net) has been collecting tweets related to the Occupy Movement since October 2011. This poster presents the challenges that we as well as other social scientists have faced when conducting social media research. These unique challenges span technical, methodological, and ethical issues that are not often foregrounded in the academic literature. We describe our progress and experiences addressing these issues and describe the development of our data collection, processing, and analysis toolkit. The SoMe Lab was recently awarded a $1 million NSF grant to develop an open source toolkit to make it feasible for non-technical researchers to conduct similar social media research.

Keywords: social media, big data, methods, open source, Twitter
Remediating Tinker Bell: Childhood Commodification and the Transmedia Narrative

Eric M. Meyers
University of British Columbia
eric.meyers@ubc.ca

Lindsey Krabbenhoft
University of British Columbia
lmkrabbenhoft@gmail.com

Julia P. McKnight
University of British Columbia
julia.p.mcknight@gmail.com

Abstract

The 100-year trajectory of the mischievous Tinker Bell, from J.M. Barrie’s 1904 play Peter Pan to the present-day Disney Fairies franchise, is a metanarrative of adaptation and remediation through which media and “childhood” can be seen to interrelate as mutually constitutive forces. With a focus on contemporary children’s media and narrative, this paper examines multiple incarnations of a media franchise at 50-year intervals. Our close-reading yields insights into the reflexive relationship between social constructions of childhood, the evolution of narrative in children’s literature, and the development of media for child audiences since the Edwardian era.

Keywords: childhood, intertextuality, material culture, digital media, remediation

Introduction

One hundred years ago, children in theatre audiences and nurseries throughout the Western hemisphere clapped to save her life. Today, 20 million children per year embody her character online as virtual fairies. The 100-year trajectory of the mischievous Tinker Bell, from J.M. Barrie’s 1904 play Peter Pan to the present-day Disney Fairies franchise, is a metanarrative of adaptation and remediation through which media and “childhood” can be seen to interrelate as mutually constitutive forces. With a focus on contemporary children’s media and narrative, this poster examines multiple incarnations of a media franchise at 50-year intervals—beginning with Barrie’s play and novelization of Peter Pan at the turn of the 20th century, Disney’s animated retelling of Peter Pan in 1953, and the Disney Fairies franchise which “rebooted” the classic character in 2005. Our close-reading of this franchise, informed by the theories of remediation (Bolter and Grusin, 1999) and intertextuality (Wilke, 1999), yields insights into the reflexive relationship between social constructions of childhood, the evolution of narrative in children’s literature, and the development of media for child audiences since the Edwardian era.

At a cultural moment in which we approach the view of literature as an interactive medium and children as content-creators, we are mindful of the ways in which commercially formulated media empower and constrain childhood agency. For younger online players, being online is becoming an increasingly important part of social life; massively multiplayer online games, or MMOGs, have become mainstream forms of social interaction (Kafai 2008, Gee 2007, Meyers 2009). Kasturi (2002) argues that in both content and form, new media conveys specific pedagogical messages to children about representation and consumption, identity and social status. In this context, the character of Tinker Bell is particularly instructive because she was developed in the pivot-point of social concern and fascination with childhood as a special time and place separate from adults, something to be both protected and commodified through integration into social policy and the economy (Gavin & Humphries 2009; McGavock 2009). In our study, we trace the practice of remediation and the gradual intensification of the child as consumer, as well as the evolution of the franchise into a true transmedia narrative (Jenkins 2003). In the process, we demystify the idea that current children’s media constellations are novel developments; rather, they fit into a historic tradition of constructing childhood around social control and commercial gain.

Acknowledgement: This work has been funded by the Social Science and Humanities Research Council of Canada, Grant #410-2011-2012.


Copyright is held by the authors.
Method and Data Source

We approach the character of Tinker Bell through a close-reading method (Bizzocchi & Tannenbaum 2011; Looy & Baetens 2003) involving engagement with several different levels of text and media to present them as a narrative of remediation: a meaningful whole. We emphasize multimodal engagement with contemporary digital media and film, while referencing and re-reading related historical “texts”, including various versions and incarnations of Barrie’s Peter Pan, Disney’s Peter Pan, and the contemporary Disney Fairies franchise. Combining textual analysis with historical/cultural analysis, we juxtapose three historical phases in the social construction of the child, the status of Tinker Bell, and the media incarnations (plays, picture books, novels, cinema, branded merchandise, games, interactive spaces) that support these constructions. The result is a metanarrative of the Tinker Bell character in which we describe a historical phenomenon: the story of how media and childhood interrelate.

Findings

Our study identifies and focuses on three phases in the evolution of the Tinker Bell transmedia franchise over a century-span. While the limits of this abstract prevent us from illustrating the extent of our findings, our final poster will lay out these 3 phases chronologically, linking the Tinker Bell character, her media instantiations, and key touchstones in the changing nature and conception of childhood. Very briefly:

Phase 1: Edwardian Era (1900s)

In the Edwardian era, “childhood” became recognized as distinct from adulthood in terms of desires, needs, and ways of being. Children were celebrated for being at once precious and charming, savage and uninhibited. This time is characterized by an unprecedented increase in children’s fiction, children’s toys and merchandise, and child welfare legislation. The play Peter Pan was introduced in this period, and became an instant transmedia success. The novelization of the play, Peter and Wendy, was sold at various price points to achieve commercial saturation, and Peter Pan merchandise was among the most popular Christmas gifts in the early 1900s.

Barrie positioned children as the link between Neverland and the “Mainland”. Fairies at the turn of the twentieth century were characterized by lawless, hedonistic, childlike behavior (Tatar 2011). Barrie’s Tinker Bell is described as buxom and flirtatious: she is also like a pre-linguistic child. The audience/reader requires Peter’s translations in order to understand what she says. Murray Pomerance (2009) describes Barrie’s Tinker Bell as representative of a particular historical moment in terms of the cultural, technological (electricity), and artistic elements that inform her relation to contemporary notions of domesticity, womanhood, and the fantastical.

Phase 2: Cold War Era (1950s)

In the 1950’s, the Disneyfication of Peter Pan intersected with a cultural moment in which society reneged on women’s war-time independence and popular culture shifted its focus away from the child alone to pervasively endorse the nuclear family unit and women’s domesticity (Pomerance 2009; Wojcik-Andrews 2000; Ohmer 2009). Of the cross-platform, multi-media merchandise surrounding Disney’s 1953 film Peter Pan, Susan Ohmer explains: “Merchandising has been a crucial part of the Disney empire since the 1920’s, but with Peter Pan it took a quantum leap forward” (2009). Peter Pan, and its attendant multi-media constellation—expanded from that of the Barrie franchise 50 years earlier through technological innovations in film animation, vinyl recordings, and the all-new immersive experience of the theme park—also intersected with a meteoric rise in material consumption. With the release of Peter Pan, Disney merchandising maximized the film’s synergistic possibilities, targeting both children and parents, creating the cultural perception of a link between youth and consumption.

Tinker Bell’s physicality in Disney’s 1953 movie Peter Pan departs from the Victorian depiction of fairies to reflect contemporary beauty ideals. Contemporary critics described Tinker Bell as a “vixen”, likening her to Marilyn Monroe, and Zsa Zsa Gabor (Pomerance 2009; Ohmer 2009). Tinker Bell’s femininity is positioned in competition with Wendy’s traditional, domestic womanhood. Like Barrie’s Peter
and *Wendy*, Disney's *Peter Pan* came about at a time of transition, and in response to contemporary attitudes and values (Ohmer 2009).

**Phase 3: Internet Era (2000s)**

Fast forward to the current media mix ecology that is the Disney Fairies franchise and we find Tinker Bell at the center of a hypermediated riot of licensed merchandise, print and electronic media, all of which point toward a subscription-access virtual world as the central, authoritative text. In 2005 Disney debuted the Disney Fairies franchise with the release of Gail Carson Levine’s novel *Fairy Dust and the Quest for the Egg*, the first installment in a trilogy aimed at 6 to 10-year-old girls. Based on the character of Tinker Bell and the limited mention of fairies and their home in Barrie’s work, Levine develops a rich, magical community centering on a small group of culturally diverse fairies lead by Tinker Bell. As a Newbery award-winning author famous for contemporizing fairytales, Levine lent credibility to the franchise, anchoring it in the cozy lap of classic children’s literature. Based on Levine’s adaptation, in 2008, Disney released the animated film *Tinker Bell*, featuring the titular character as fully developed with a voice and a relatable backstory that locates her within a social structure that is recognizable to children as the class system of our own society.

Disney concurrently released the Pixie Hollow MMOG, which approximately 20 million children visit per year. While watching the movie is not necessary, it functions as a guide that brands the experience of playing Pixie Hollow by contextualizing the games within a cross-platform narrative structure. In the MMOG, children create their own embodied fairy based on Tinker Bell and her fairy friends; 40.5 million have been created to date. Narrative elements of the films are mined and incorporated into the Pixie Hollow MMOG, illustrated books, licensed toys and other merchandise, a magazine, a theme park and Ice Shows to form a vast narrative that crosses a multitude of platforms.

Children's media franchises in the current era bank on a the consumption of interrelated media enhancing the richness of the ‘reading’ experience, and position the serial/subscription environment in the center of a media constellation that aims for integration into “the routines of children’s everyday lives” (Buckingham & Sefton-Green 2004; see also Ito 2008). Scholars suggest that virtual worlds offer children an ontological position not available to them in children’s literature, or anywhere in their lives, granting them agency and authority. Are virtual worlds a reflection of or a force in the (re)construction of contemporary childhood (Papazian 2010; see also Sherington 2010)? We argue that they are both.

**Scholarly Contribution**

The historical and cultural analysis of a century-long transmedia narrative stands out from other work that focuses only on contemporary instantiations of transmedia. By charting the evolution of a popular character over several periods of remediation, we gain a richer perspective on the relationship between media and childhood, a perspective largely absent in media franchises that emerge (and sometimes disappear) in less than the span of a single generation. Our work illustrates the changing nature of character and narrative in children’s media, but also shows how many “commercial transmedia supersystems” (Kinder 1991), portrayed by some as novel aspects of the digital age, are rooted in a century-long commercialization of the childhood experience.

A novel aspect of the transmedia narrative that emerges from this study is the shifting focus of the seminal text toward the immersive, embodied child experience. In the Internet Era we see the rise of the virtual world as a centerpiece of the narrative, surrounded by supporting media (movies, novels, comics, websites and fan spaces, etc.) that might have been centerpieces themselves in a different age. These new central texts are dynamic and immersive; at the same time they are highly stratified and consumption-focused. While this move might appear to facilitate greater child agency and a different level of engagement with the franchise, through our play in these spaces we find that the commercial constraints counter what appears on the surface to be an empowering shift. Our analysis of the changing nature of the transmedia narrative, situated in a historical understanding of children’s media and culture and with a spotlight on an iconic media symbol, will be of interest to scholars of new media, educators of the ‘Net Generation, and the designers of contemporary media experiences for children.
References


Trust and Community: Continued Engagement in Second Life

Peyina Lin
pl3@uw.edu
Natascha Karlova
nkarlova@uw.edu
John Marino
marinoj@uw.edu
Michael Eisenberg
mbe@uw.edu

Information School, University of Washington

Abstract

3D social virtual worlds (VWs), like Second Life (SL), have potential as alternative modes for information seeking and socialization. Yet, finding trustworthy social connections for expertise and social support can be challenging in Second Life (SL) where identities are fluid. We explored how long-time SL users established and diffused trust within their communities by vetting each other as trusted information sources. By drawing on the relationship among trust, being highly sought after, social status, and continued engagement in SL, we identify challenges that VWs like SL face, and provide recommendations for social navigation systems to support continued engagement in social virtual worlds.

Keywords: virtual world, trust, community, social navigation, continued engagement

Introduction

During our 180+ hours of exploration in Second Life® a 3D social virtual world (see box 1), we learned that despite media coverage (Hansen, 2009) about decreased activity in Second Life (SL), a set of long-time users remained highly involved. They contributed key content to SL and were central to sustaining their virtual communities. Belonging to marginalized communities (e.g., multiple sclerosis, transgender, alternative political groups) was one reason these users were continuously involved in SL. For example, those from communities with scarce members more easily found other world-wide members in SL. However, such community membership did not fully explain the continued engagement in SL by a set of long-time SL users. In looking for trustworthy information sources during our exploration, word-of-mouth led us to the same set of long-time users, who were regularly being referred to and sought after for expertise. This led to our research questions:

• How does the interaction among trust, being sought after, and social status contribute to the continued engagement of Second Life (SL) by long-time users?
• How can the design of social navigation systems support virtual world community practices while improving on the challenges that existing practices raise?

These are important research problems. First, 3D social virtual worlds (VWs) have potential as alternative modes of socialization. Yet, the widespread use of social 3D VWs has been slow. Thus, understanding the continued engagement in SL by long-time users can contribute to helping more people from diverse populations find community via virtual worlds. Second, social navigation systems are increasingly important as more people use online tools to socialize, seek trusted domain experts, or meet their activity needs. By understanding the challenges brought forth by the interaction among trust, being sought after, status, and continued engagement in SL, we provide conceptual recommendations for how social navigation systems can better support the continued use and engagement in 3D social VWs like SL.

Acknowledgement: John D. and Catherine T. MacArthur Foundation, grant No: 92258-0; research participants.
Copyright is held by the author/owner(s).
Related Literature

The information research community has been interested in the use of social networks for information seeking (Morris, Teevan & Panovich, 2010) or community membership (Ostrander, 2008), and in ways to improve such experiences through social navigation systems (Terveen & McDonald, 2005). Crucial to the process of assessing information from social networks is trust, or the commitment to an action which depends on the behavior of the trusted party (Golbeck, 2009); e.g., choosing to use the information provided by the trusted party to inform other actions. Trust has been studied in the context of profiles in online social networks (Golbeck). Yet, few have investigated how social navigation systems may help diffuse trust in 3D social virtual worlds like SL, where users can keep their first-life identities private, and in ways that address the sustainability of the virtual community. The research community has also been interested in what contributes to the continuous use of various technologies (Joinson, 2008; Lampe, Wash, Velasquez & Ozkaya, 2010), yet little exists on the continued engagement in VW’s like SL. Our study fills these gaps.

Method

We transcribed 10 in-depth semi-structured interviews (120 mins. each) with long-time SL users. We took observation field notes of day-to-day activities over critically-sampled 5 hours, per participant, across 2 weeks. We used the “sensitizing” approach to guide inquiry and analysis, to allow the “immediate world of social experience… to shape and modify” our understandings of the research (Patton, 2002, p.278). This enabled us to view participants holistically, embedded in socio-structural contexts that contributed to continued engagement. Three researchers cross-validated their data analyses using a constant-comparative method, iteratively extracting and validating emerging patterns in the data.

Participants were highly involved long-time users. Most of them contributed valuable content to their communities: producing live shows, creating educational materials, acting as information resources, and providing community support.

Findings

Predispositions Contributing to Continued Engagement

Diverse personal and social predispositions emerged as the factors contributing to the continued engagement in SL by long-time users (see Table 1). Predispositions refer to the conditions which define each user before being exposed to stimuli that can potentially change these conditions. In Table 1, after “post-initial use”, the ways in which long-time users gained status leading to their increased engagement
and continued use stood out. For example, a member of a technology community is more likely to hear about SL and be an early user (Initial Use, Social Predisposition in Table 1). After exploring SL for its novelty, long-time users were sought after for advice on their area of expertise either because of the description in their SL profile or the trust established via interpersonal interactions (Post-initial Use, Personal Predisposition). The trustworthiness of these experts is spread throughout the virtual community through word-of-mouth, i.e., recommendations from those who consulted such expertise. As a result, these experts are more likely to be sought after for advice (Post-Initial Use, Social Predisposition). Consequently, they produce more content to help others and in the process develop an increased sense of community, both of which lead them to further increase their level of engagement in SL (Continued Engagement, Personal and Social Predisposition).

Table 1

<table>
<thead>
<tr>
<th>Use-stages</th>
<th>Predispositions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Personal</td>
</tr>
<tr>
<td>Initial Use</td>
<td>• background: former gamer</td>
</tr>
<tr>
<td></td>
<td>• curiosity, fun, novelty</td>
</tr>
<tr>
<td></td>
<td>• physical constraint: homebound</td>
</tr>
<tr>
<td></td>
<td>• searching new VWs</td>
</tr>
<tr>
<td>Post-initial Use</td>
<td>• expertise in profile</td>
</tr>
<tr>
<td></td>
<td>• vested interest (exploring social identity, pleasure of super physical power)</td>
</tr>
<tr>
<td></td>
<td>• personally valued mission in SL (raising awareness on abuse)</td>
</tr>
<tr>
<td></td>
<td>• honest practices</td>
</tr>
<tr>
<td>Continued Engagement</td>
<td>• personal gains (social identity, extend FL profession, e.g., therapist)</td>
</tr>
<tr>
<td></td>
<td>• ability to contribute</td>
</tr>
<tr>
<td></td>
<td>• learning</td>
</tr>
<tr>
<td></td>
<td>• contribute to social welfare (e.g., amnesty)</td>
</tr>
<tr>
<td></td>
<td>Social</td>
</tr>
<tr>
<td></td>
<td>• member of tech community</td>
</tr>
<tr>
<td></td>
<td>• recommended by ties</td>
</tr>
<tr>
<td></td>
<td>• socially disenfranchised: seeking to socialize</td>
</tr>
<tr>
<td></td>
<td>• word-of-mouth spread of expertise → trustworthiness</td>
</tr>
<tr>
<td></td>
<td>• asked for expertise</td>
</tr>
<tr>
<td></td>
<td>• being recommended and sought after repeatedly → high status</td>
</tr>
<tr>
<td></td>
<td>• publicity attracting attention</td>
</tr>
<tr>
<td></td>
<td>• community or sense of community:</td>
</tr>
<tr>
<td></td>
<td>• tightly knit community</td>
</tr>
<tr>
<td></td>
<td>• easy access to community resources</td>
</tr>
</tbody>
</table>

A First Life Identity is Not Necessary

Importantly, for long-time SL users with some domain-expertise, being highly sought after did not necessarily depend on linking their first-life identities or expertise to their SL profile information. For example, Eni kept his first life and SL identities separate, and yet was a trusted activist from whom others sought advice: “They wanted to know if I could help them out organizing this [virtual protest].” There were others who thought a connection to first life was not needed to assess trust: “I don’t know what their name is, what their real life name is. That’s okay.” (Eva). For those who did not connect their SL avatars to their first life identities, being sought after depended on the trustworthiness they established through interpersonal interactions and on social recommendations. For example, Win states: “I think we trust people based on our entire experience with them… with M, we’re at a point where anything she said to me I would take at face value… and that’s based on several years of working together… a lot of alignment… I think that’s what creates trust.” The interaction between trustworthiness, being recommended/ highly sought after, and status, in increasing degree over time, was crucial to enabling continued engagement (see fig. 2).
Community Challenges Faced by Relying on Word-of-Mouth

Word-of-mouth, as the typical mode of finding expertise, can create various problems for a virtual world community. First, although vetting each other as trusted information sources in their communities via word-of-mouth contributed to tightly-knit communities and enabled easy access to community resources, such repeated referrals amongst long-time users may threaten the community’s longevity. In the long run, repeat-referrals to the same few puts undue burden on a few long-time users, and disengages them from others they cannot attend to in a timely manner. Second, with repeated referrals to the same few, other less well-known community members with expertise may feel marginalized or undervalued and thus are more likely to leave the community.

Recommendations

Since being highly sought after plays a key role in continued engagement, it is critical for social VWs to support social navigation adequately while not jeopardizing the longevity of the community. Thus, a social navigation system in SL should allow leaders to mitigate or share the burden of leadership and allow community members to broadcast their own expertise. In addition, since interpersonal interactions were critical to diffusing trustworthiness via word-of-mouth, and in increasing our long-time users’ level of engagement, a social navigation system in SL would need to provide a way to diffuse trustworthiness in a personal way—with a ‘personal touch’. For example, long-time users could bestow status on others by providing a publically viewable endorsement in the system (vetting other community members they trust). Such an endorsement showcases the expertise of other community members, shares the burden of being constantly called upon for expertise, and then the system matches the vetted community members with expertise seekers (see fig. 3). Importantly, a ‘personal touch’ can be added to these endorsements by enabling expertise seekers to verify such endorsement, in situ, through interactions with others they trust.
Conclusions

With this study, we contributed in the following ways. We identified unique sets of personal and social predispositions at different use-stages which contributed to long-time users’ continued engagement in SL. We described how the interaction among trust, being sought after, and social status, in increasing degree over time, contributed to the continued engagement. We discussed the community longevity challenges faced when long-time users vet each other as the primary trusted information sources, and provided recommendations to support continued engagement by a broader base of users in the community (rather than the same set of long-time users). Our recommendation capitalizes on the importance of interpersonal trust in this community by offering long-time users the opportunity to vet other community members. This differs from other expert recommending systems (for a thorough review, see Terveen & McDonald, 2005). Because ‘personal touch’ is irreplaceable by an automated system, the recommendations aren’t fully system generated but go through a filtering stage by well-trusted community members. Prolonged engagement in data collection showed strong evidence that social validation from trusted community members was particularly important in SL, where identity is fluid and often anonymous. Our conceptual recommendation integrates aspects of that ‘personal touch’ with social matching for SL.

References


New Models for Supporting Digital Scholarship: UNT’s DiSCo

Anjum Najmi  
University of North Texas  
Digital Scholarship Co-Operative  
an0091@unt.edu

Spencer D. C. Keralis  
University of North Texas  
Digital Scholarship Co-Operative  
spencer.keralis@unt.edu

Abstract

This poster examines a new model for supporting digital scholarship across the disciplines, as implemented by the University of North Texas Digital Scholarship Co-Operative (DiSCo). Digital scholarship offers researchers ways to communicate and collaborate across disciplinary boundaries and provide opportunities for training and access to digital tools to support research and teaching. Recognizing this needs, UNT established the DiSCo to serve as a focal point for data management support, and to foster the development and use of digital tools for research, teaching, and learning.

Keywords: digital scholarship, interdisciplinary research, collaboration, community-based

Introduction

This poster presentation examines digital scholarship and its impact on the humanities and interdisciplinary education. It explores new models for scholarship, and touches on the cultural implications for research, communication, and knowledge generation. It describes the implementation of a co-operative model of digital scholarship support implemented at the University of North Texas. This model is research-based, and intended to intervene in a range of disciplines. The Boyer Commission Report (1990) advocated for the scholarship of teaching, learning and service as key components to be valued by educational institutions and government policy. As technology use and the digitization of material continue to increase, universities and research institutions find themselves facing a paradigm shift (Feinberg, 2004). Technology offers new possibilities instigating change, compelling institutions to rethink and adopt new methods of scholarship (Lui & Thomas, 2012). UNT’s DiSCo is one model for supporting digital scholarship in a rapidly changing academic environment.

Background

Digital scholarship has the potential to fundamentally transform the way research, teaching, and learning occur in higher education. Traditionally, the function of the university or professoriate has been to generate knowledge. Scholarship is viewed as an individualistic endeavor with scholarly practices varying across disciplines. The natural sciences follow a data-centric approach that consists of asking clearly defined questions, collecting and analyzing data, using standardized methodology and potentially obtaining replicable and generalizable results. Large scale, capital-intensive collaborations support individual scholars in the scholarly process. Distributed groups and teams achieve a greater degree of specialization and division of labor (Galison & Hevly, 1992).

In the humanities, the focus has traditionally been on selecting and synthesizing sources around common themes, to note subtleties that cause one artifact to be different from another. It is about "investigating the meaning and form of a concept” and is dependent on the researcher’s interpretation, reflection, and evaluation (Feinberg, 2004 p. 21). Typically funding was limited, and emphasized individual awards to scholars, and projects existed within disciplinary boundaries with teams pursuing their own individual objectives with limited meaningful interaction with one another (Smith, 2003; Nichols, 2009).

Acknowledgements: The UNT Digital Scholarship Co-Operative was established through the generous support of the UNT Libraries and the College of Arts & Sciences.


Copyright is held by the authors.
Boyer in his seminal report proposes a more inclusive view of scholarship "a recognition that knowledge is acquired through research, synthesis, practice, and teaching" (Boyer, 1990, p. 24). This includes four dimensions; discovery: original research to advance knowledge in a specific discipline, integration: extending knowledge across disciplines and time, application: engaging a wider audience beyond the field to extend the disciplinary knowledge along with peer feedback, and teaching: the systematic study of teaching and learning with application, reflection and evaluation ingrained (Boyer, 1990; Pierce, Weller, Scanlon & Kinsley, 2010). In keeping with these principles, new technology and digital resources offer possibilities for bringing together the broad range of scholarly practices to support interdisciplinary research and scholarship (Schreibman, Mandell, & Olson; 2011). Both, universities and funding institutions are demonstrating their recognition of the need to provide support for digital scholarship.

Digital Scholarship

The American Council of Learned Societies (ACLS) defines digital scholarship as "any element of knowledge or art that is created, produced, analyzed, distributed and/or displayed in a digital medium for the purpose of research or teaching" (2007). Other experts see it as a way to support open access to scholarly knowledge and a way to promote the use of innovative research methodologies (Spiro, 2007; Lui & Thomas, 2012). For institutions, digital scholarship helps challenge traditional standards for tenure and promotion (Fitzpatrick, 2011). The past two decades have seen the emergence of numerous centers for Digital Scholarship that support interdisciplinary research in different ways. Some exist within single institutions, supporting diverse projects, programs and activities, while others support multi-tiered projects across institutions. The centers bring together funding sources and a mixture of professional expertise from K-12 educators, higher education, government and industry (Bos et. al., 2007, Nowviskie, 2011). Notable examples of center-based projects are The Valley of Shadows, Rome Reborn (Institute for Advanced Technology in the Humanities, University of Virginia) and ventures such as Project Bamboo (Zorich, 2008, Mellon Foundation, 2010)

Models for Support

There are three principle models for supporting digital scholarship, the enterprise-based model, the innovations model, and the community-based model.

On the enterprise end of the spectrum is the Roy Rosenzweig Center for History and New Media (CHNM). CHNM’s focus is on tools and projects for digital humanities research. CHNM makes historical collections available online and develops tools for researchers, like Zotero for collecting, managing and citing research sources, and Omeka, a platform for online exhibition. Websites like Teachinghistory.org and History Matters provide secondary educators with teaching and learning resources. CHNM established THatCamp, the unconference model for collaborative discourse on digital humanities topics. Participants propose their own topics and follow workshops to demonstrate particular tools and methods. Attendees include scholars, students, librarians, archivists and museum professionals. CHNM follows the enterprise-based model where funding is fixed, ongoing and regulated and sponsors are government institutions and private foundations such as the Department of Education, the Library of Congress, the Institute of Museum and Library Services, the American Council of Learned Societies, and the Mellon, Sloan, Rockefeller, and Kellogg foundations among others.

An example of the innovations model of scholarship is The Brooklyn Institute for Social Research. The Institute maintains an online presence where materials, intellectual support, and space are offered to young scholars to teach write, research and publish. Social media such as a blogs and Twitter give members a voice, connect them with each other and help them to stay informed. A platform for creative projects allows people to post their project ideas, make bids, pledge support, earn funding, and share rewards, using a crowd-sourced funding model to support individual projects. Examples of successful projects are Diaspora: an open source personal web server, and Explore Newport: a 3D virtual tour of the historic town of Newport. Community building is central, funding a shared responsibility, peer support and reward embedded in the process.

The Digital Scholarship Co-Operative (DiSCo) at the University of North Texas follows a community-based model of scholarship, customizing features based on the needs of its constituents and then aligning institutional goals to reflect these values. DiSCo supports graduate students and faculty in
integrating digital research tools into their teaching and learning by providing resources for student projects, born digital art and collaborative works. It will offer fellowship and employment opportunities for graduate and undergraduate students to develop skills in digital tools, project management, collaboration and research. Evaluation of data management and digital curation practices help develop best-practice recommendations, and are part of the institution-wide infrastructure. The Co-Op seeks to champion Open Access at all levels of scholarship.

DiSCo supports its community by connecting users and their research needs to centers of excellence and expertise within the university without duplicating extant services. The Co-Op views itself as a router, receiving requests from scholars for resources, and forwarding the researcher to extant service centers. Where unmet needs exist, the Co-Op can evolve in response to fill gaps in the research needs of its stakeholders.

Digital resources provide the platform for collaboration, communication and research. With these objectives in mind, the Co-Op sponsors periodic webinars and workshops with national leaders in digital scholarship, hosts the annual Digital Frontiers conference, established an inter-institutional humanities colloquium to facilitate trans-disciplinary conversations between technologists and scholars, and participates in International Open Access week programming. The goal is to foster the creative use of digital resources in research, teaching and learning through an environment of coordination and collaboration across campus with DiSCo as its hub. This has the potential to minimize minimizing the siloing of skills and expertise. These initiatives come together to serve the University’s Four Bold Goals (http://www.unt.edu/features/four-bold-goals/), helping fulfill the institution’s broader educational mission.

Figure 1 - DiSCo & UNT’s Four Bold Goals

Conclusion

Digital scholarship offers researchers new ways to communicate and collaborate across disciplinary boundaries, allowing for more democratic means of knowledge creation. As centers continue to advocate for scholarship the concept does pose certain challenges, with collaborative efforts often translating into the sharing of resources more than the sharing of knowledge, and give rise to redundancies in efforts as centers inadvertently become holders of information instead of disseminators of knowledge (Zorich, 2008). The community-based Co-Operative model advanced by UNT seeks to alleviate these redundancies by facilitating access to existing services in response to the needs of its
stakeholders. Further detail and in depth aspects of digital scholarship will be discussed at the conference poster presentation.

References


Feinberg, Melanie. (2012). Information studies, the Humanities, and Design research: Interdisciplinary opportunities. iConference 2012, Toronto, Ontario, Canada

Four Bold Goals for the Future of UNT: University of North Texas Strategic Plan 2012-17 http://www.unt.edu/features/four-bold-goals/


Roy Rosenzweig Center for History and New Media (CHNM). George Mason University, Fairfax, Virginia. http://chnm.gmu.edu/


Spontaneous Reactions to an Anti-piracy Initiative: A Youtube Clip Micro Analysis

Gustaf Nelhans  
gustaf.nelhans@hb.se  
Hanna Maurin Söderholm  
hanna.maurin@hb.se  
Jan Nolin  
jan.nolin@hb.se  
Mathias Klang  
mathias.klang@hb.se  
Monica Lassi  
monica.lassi@hb.se

University of Borås, Sweden

Abstract

In this case study we analyzed the traces of spontaneous reactions of Youtube users when confronted with the short clip 'You wouldn't Steal a Car', that was used by the Motion Picture Association of America (MPAA) to influence people not to download copyrighted material from the Internet. This film has become an important cultural icon, which to a certain degree has shaped a whole generation of film viewers. The aim of this study was to provide an example of how anti piracy initiatives are received and understood by the receivers of the message. This was performed by collecting and analyzing the users spontaneous reactions as entered as comments on the Youtube page for the clip by qualitatively categorizing the contents using a bottom up approach. The results suggest that people practicing Internet-based culture consumption (IBCC) do this in more nuanced ways than is assumed in the film, where they are polarized as either "common thieves" or "good citizens".

Keywords: Internet-based culture consumption, spontaneous reactions, anti-piracy, Internet

Introduction

The MPAA is a U.S. trade organization for six major Hollywood Studios, promoting their business interests. As part of that it has taken a strict stance on copyright infringements fighting the sharing of copyrighted material on the Internet, focusing on efforts involving influencing the practices of young people. The clip featured in this study was part of an advertisement campaign introduced in 2004. It appeared as a pre-movie advertisement in theaters, or included in DVDs without the possibility to skip before program content. This clip has been discussed in scholarly literature in relation to the legal use of the term 'theft' in intellectual property discourse (Loughlan, 2008) and as an example of both the "confuzzling" arguments used by stakeholders, (Yu, 2011) as well as the strategies used by representatives of record companies and artists in the struggle of defining stakes in the controversy of file sharing. (Martin, Moore & Salter, 2010) Youtube comments in general was studied by Lange (2007) however, viewers’ actual reactions to this specific clip have not been discussed other than in general terms, often by the researchers themselves when reflecting on the paradox that the clip is only shown to viewers paying for a DVD or a ticket to the movie, and seldom included on movies that have been pirated from the original source.


Copyright is held by the authors.
The goal of this ongoing study is to capture and characterize these reactions and contribute to the emerging field of research about IBCC practices (Nolin, 2010) that avoids taking the usual clichés of polarization into legal or illegal, or, good or bad, conveyed by the anti-piracy organizations for granted. This work is part of a larger research project in which we aim for a symmetrical view towards the various forms of sharing of media, including both legal and illegal use.

Methods

This study focuses on viewers' reactions to this clip as expressed in the comments on its Youtube page. We argue that studying comments in a web forum or other web media in reaction to a specific event (in this case a video clip) provides methodological possibilities to identify situational and immediate reactions of users that would otherwise have been hard to capture. These “micro scripts” or “nano texts” (Cronin, 2012) produced by the viewers, seldom contain more than a few words or sentences, yet they provide us with what seems to be rather instantaneous reactions of a kind that is hard to acquire by using more conventional methods. Additionally, traditional methods of analyzing reactions as those identified above entail methodological challenges, including questions about representativity, and, the infamous “researcher effect”, which although challenging to distinguish from other influences presents itself as one of the methodological horrors of social science. (Woolgar, 1988) Here, only viewers that were logged in to the Youtube site were able to comment on the clip.

Our data consists of all the posted comments and pertain mostly to direct reactions to the film’s content or the producers of it, and exhibits low networking level in terms of discussions and interactions between people. Therefore we chose a straightforward content analysis (e.g. Neuendorf, 2001) using two sets of data. First, the metadata information regarding the clip and comments received on the Youtube page was collected. This includes viewer statistics and usage data that functions as quantitative traces of the viewers’ actions. Second, we conducted a qualitative content analysis, focusing on the contents of the comments using a bottom up approach where the coding scheme emerged during the analysis. The coding was done in two steps: first we identified three general categories corresponding to judicial, emotional, or content-based issues; followed by a secondary categorization to classify additional nuances within the code categories.

The Movie Clip: You Wouldn’t Steal a Car

This section presents a general description of visual characteristics, content, and structure of the movie clip. The clip is cut fast, using suggestive imagery depicting various illegal actions accompanied by very aggressive music. Visually, it is presented in a “silent movie” style, interfoliating imagery and text. This resembles the narrative technique of Sergei Eisenstein, with a structure including thesis – antithesis resulting in a synthesis (Figure 1.). The narrative consists of:

YOU WOULDN’T STEAL A CAR
YOU WOULDN’T STEAL A HANDBAG YOU WOULDN’T STEAL A MOVIE. DOWNLOADING PIRATED FILMS IS STEALING, PIRACY IT’S A CRIME.
REPORT IT.
1 800 251 996
WWW.MOVIEPIRACY.ORG.AU

The crimes used to depict unlawful behavior are largely simple crimes, ranging from stealing a car and a handbag—actions that are easy to call illegal—to stealing a DVD, with the aim to demonstrate how downloading a film is a similar act of crime. The clip also contains cultural references to other media and films that are known to the target audience of the campaign. For example, the text is seemingly modeled after the film Se7en, which used the shaky haphazardly shown text to provide a sense of haste, urgency and aggressiveness.
Figure 1. Screenshot storyboard
Results

Audience Details

The clip is available in many Youtube versions. We chose ‘Movie Piracy - It’s a crime, by the
Australian government’\(^1\), based on it having the highest viewer count (228,051 views as of May 24, 2012)
and being the oldest version available on the site, uploaded in May 2006.\(^2\) In all 1283 comments have
been added to the Youtube page that were used as a source for the analysis. The primary visiting
audience is males in the age interval 18-44. Fig. 2 illustrates viewers’ geographical distribution and
relative popularity (darker: more popular, lighter: less):

![Figure 2. Viewers’ geographical distribution](image)

Audience Reactions

The audience reactions pertained to judicial, emotional and content related issues. This section
presents some illustrating examples of our findings. Many comments argued that there was a difference
between stealing a physical object and a digital copy, because the digital original would still be available
to the owner. One of the most frequent comments simply stated that:

“I would if I could”

Counter assertions were given in various degrees of meaningfulness:

"you wouldn't download 'a car'?a hand bag'?a bear'?...?'a girl'?CHUCK NORRIS"

On the other hand, a large part of the comments stated in various degrees of certainty that they
would commit all of the crimes depicted.

Emotional comments ranged from meaningless slander to mere personal statements. In other
cases commenters argued for how and why they download:

“I download games & movies to test that they are worth of buying. I prefer to support developers if
the game I downloaded (or movie ) is good”

---

\(^1\) http://www.youtube.com/watch?v=l5SmrHINWhak
\(^2\) The “Australian Federation Against Copyright Theft” produced this version. Music and imagery is largely the same as in the
American version, except that one film section is cut and some of the imagery is even faster paced. In the end, a banner referring to
AFA©T, The Australian Federation Against Copyright Theft is shown. The referred web page is inactive, but accessible via Wayback
sets some time frame limits for the campaign.
A large share of the comments was directed towards the content. These ranged from issues regarding various technologies used in the film, the fast Internet connection the girl seemingly had, or specific elements used in the film, e.g. text effects and fonts. The music attracted a great deal of comments, mostly favorable to the sound, requests where to “get” it, and who the artist was. Many stated that the artist was the band Prodigy, while concerns regarding the copyright status of the music were raised, e.g.:

“No Man Army by the Prodigy, wonder if they got permission or just knocked it off”
Another stated that it
“would be funny if whoever made this advert didn’t have the rights to use this song”
Interestingly, the music was allegedly used by the MPAA without permission. On December 1, 2011, the site Torrent Freak published an article describing how the Netherlands composer Melchior Rietvelt, who composed the music for a local film festival, had found his music used in the clip by MPAA in a setting allegedly not covered by the licence it was originally agreed. (Ernesto, 2012)

Conclusion

The methodological claim of this study is that comments on a Youtube clip can be viewed as spontaneous reactions to an event that provides access to peoples’ reactions, which to some degree was illustrated by the results.

The results provide resources for arguing that people practicing IBCC do this in more nuanced ways than is assumed in the film, where the MPAA tries to polarize users as either “common thieves” or “good citizens”. Especially, we found that downloading of copyrighted material by itself is not stated to preclude users to purchasing the same or other material.

References


CHAOS: User-driven Development of a Metadata Scheme for Radio Broadcast Archives

Haakon Lund  
Royal School of Library and Information Science, Copenhagen  
hl@iva.dk

Toine Bogers  
Royal School of Library and Information Science, Copenhagen  
tb@iva.dk

Birger Larsen  
Royal School of Library and Information Science, Copenhagen  
bl@iva.dk

Marianne Lykke  
University of Aalborg  
Department of Communication and Psychology  
mlykke@hum.aau.dk

Abstract

CHAOS\(^1\) (Cultural Heritage Archive Open System) is a digital platform for Danish radio broadcasts. Radio broadcasts are an important and vibrant part of our cultural heritage, but providing efficient and effective access to such archives is challenging for lack of a solid digital infrastructure. The Danish LARM project aims to meet this challenge by making one million hours of radio programs available to humanities researchers through the digital platform CHAOS. CHAOS is being built in close cooperation with the researchers involved in LARM. In this paper, we present the user-driven development of the multi-tiered metadata scheme used in CHAOS.

Keywords: metadata, cultural heritage, user-driven development, radio, digital humanities

Introduction

The past two decades have seen mass investments in large-scale digitization projects aimed at digitally preserving our cultural heritage. In more recent years, the focus of such efforts has turned to the question of how to provide efficient and effective access to these digitized collections for both academia and the general public. One such project is the Danish LARM project\(^2\), which is a joint initiative between the Danish national broadcasting corporation (DR)\(^3\), the State and University Library (SB)\(^4\) hosting the Danish Media Archive, and a consortium of Danish university humanities departments.

The goal of LARM is to unlock the true potential of Danish digitized radio broadcast archives by providing access to over a million hours of radio to humanities researchers through a dedicated digital infrastructure called CHAOS. CHAOS offers streaming access to all the material broadcast in the period 1985-2005 and selected broadcasts dating all the way back to 1925. Examples of the diverse range of use cases the project caters to, are investigating the changes in the Danish language as used on the radio, and analyzing the changing rhythm and pace of radio broadcasts throughout the years.

---

3 [http://www.dr.dk](http://www.dr.dk)
4 [http://en-statsbiblioteket.dk/about-the-library](http://en-statsbiblioteket.dk/about-the-library)

Acknowledgements: The LARM-project is supported by The National Programme for Research Infrastructure (Grant number 09-067292). We thank Assistant Professor Mette Skov, Royal School of Library and Information Science, for fruitful discussion in developing the meta data scheme.


Copyright is held by the authors.
An integral part of enabling efficient and effective access to these radio broadcasts is developing a metadata scheme for handling the archival data related to these broadcasts that can both capture the original metadata and handle the diverse user needs of the participating researchers. In this paper we present work on the user-driven development of a metadata scheme that can satisfy the needs of all the LARM partners and unlock the true potential of Danish radio archives.

User Requirements

Development of the CHAOS platform and infrastructure is characterized by a bottom-up approach where development takes place in close cooperation with the researchers taking part in the LARM project. This user-driven approach has also been an essential part of developing the metadata scheme used in LARM, which has the effective and efficient retrieval of broadcasts as its main objective. DR’s own broadcast archives contain archival metadata assigned at the level of individual broadcasts. Much of this core metadata has been transferred to the CHAOS archives, although some metadata fields could not be transferred due to copyright and privacy concerns. Therefore core metadata in CHAOS is sparsely populated, with a positive bias towards later years, and does not allow for the effective retrieval of radio content.

To optimally support the humanities researchers in their work, we performed a user requirements analysis to determine how this metadata scheme should be extended. We performed this requirements analysis through a series of workshops with the participating researchers and through an online survey. The main purpose of these workshops was to gauge the researchers’ expectations with regard to descriptive archival metadata, and to identify which types of additional annotations researchers deemed necessary to support them in their daily research practice. We found that researchers expressed a desire for not only metadata that supports the effective retrieval of radio broadcasts, but also for adding research-specific annotations at both the broadcast level as well as at segments of broadcasts. A main outcome of the requirements analysis is that the needs of humanities researchers are so diverse that it is unlikely that a single unified metadata list will suit all. As a consequence, our proposed metadata scheme includes project-specific metadata, allowing individual researchers and project to adapt metadata lists to their needs (Skov & Lykke, 2012).

Conceptual Metadata Scheme

Our main objective in creating the CHAOS metadata schema was to develop a metadata scheme that was easy to work with, easily extensible, and would provide for flexible data exchange. For this reason, the CHAOS metadata scheme was built on top of the Dublin Core Metadata Element Set (Dublin Core Metadata Initiative, 2012) as implemented by the European Broadcasting Union (EBU, 2011). Using Dublin Core as our foundation also simplifies data exchange with other cultural heritage repositories, such as Europeana (Europeana, 2012).

Our user requirements analysis revealed the following three levels of metadata relevant to LARM researchers: (1) core archival metadata, (2) LARM metadata, and (3) project-specific metadata. For each of the three levels of descriptive metadata a number of administrative metadata were identified. In addition, researchers expressed a desire for functionality in CHAOS that could support them in their annotation. Figure 1 illustrates the resulting metadata architecture developed for CHAOS based on these requirements. We describe each of these metadata levels and the annotation support systems in more detail in the rest of this section.
Core Archival Metadata

*Core archival metadata* covers metadata inherited from the original data source—usually DR or the State and University Library. This core metadata is immutably tied to each broadcast and is regarded as historical data, with all the possible flaws this might contain. Core metadata is assigned at the level of individual broadcasts and is intrinsic to the broadcast. The amount of data is limited to technical information for the most part, with a few descriptive additions, e.g., program title, producer, and a short narrative. Below is an example of some core archival metadata elements:

<table>
<thead>
<tr>
<th>Metadata element</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radio channel</td>
<td>Broadcast channel</td>
<td>P2</td>
</tr>
<tr>
<td>Program title</td>
<td>Original title</td>
<td><em>Til Italien!</em></td>
</tr>
<tr>
<td>Program start (time)</td>
<td>Date and time for beginning of program</td>
<td>18. mar. 2006 kl 19:00:00</td>
</tr>
<tr>
<td>Program end (time)</td>
<td>Date and time for end of program</td>
<td>18. mar. 2006 kl 19:30:00</td>
</tr>
<tr>
<td>ID of origin</td>
<td>Unique Production ID</td>
<td></td>
</tr>
<tr>
<td>Creator and role</td>
<td>Producer, etc.</td>
<td>Hans Hansen (Producer)</td>
</tr>
</tbody>
</table>

LARM Metadata

*LARM metadata* is descriptive metadata with the aim of enriching the sparse core metadata with more detailed information about content and participants of a broadcast. LARM metadata is also assigned at the level of individual broadcasts and is editable by all LARM members. Typically, the first researcher
to use the radio program for research purposes adds this metadata. Below is an example of some LARM metadata elements:

<table>
<thead>
<tr>
<th>Metadata element</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program title</td>
<td>If title at the archival level is absent or incomplete.</td>
<td></td>
</tr>
<tr>
<td>Person - participant</td>
<td>From help system (controlled)</td>
<td>Roger Norrington</td>
</tr>
<tr>
<td>Person - subject</td>
<td>From help system (controlled)</td>
<td>Felix Mendelsohn Bartholdy</td>
</tr>
<tr>
<td>Genre</td>
<td>From help system (controlled)</td>
<td>Koncertoptagelse</td>
</tr>
<tr>
<td>Related objects</td>
<td>Webpage, podcast, photo, etc.</td>
<td>URL to ressource</td>
</tr>
<tr>
<td>Subject</td>
<td>From help system (controlled)</td>
<td>Klassisk musik</td>
</tr>
<tr>
<td>Tag</td>
<td>User defined keyword (uncontrolled)</td>
<td></td>
</tr>
<tr>
<td>Annotation</td>
<td>Annotation related to entire show</td>
<td></td>
</tr>
</tbody>
</table>

**Project-specific Metadata**

*Project-specific metadata* is of a more analytical nature and is associated with one or more individual research projects within the LARM initiative. Project-specific metadata can describe an entire show or parts of a show. Project metadata is assigned by the researchers working on a LARM research project, and is owned by that research project. Metadata at the project level is designed to be open and flexible, but the following metadata elements are suggested:

<table>
<thead>
<tr>
<th>Metadata element</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Object title</td>
<td>Introduktion til koncerten</td>
</tr>
<tr>
<td>Person - participant</td>
<td>Person participating in show</td>
<td>Magnus Møller</td>
</tr>
<tr>
<td>Person - subject</td>
<td>Person as subject for the show or part of show</td>
<td></td>
</tr>
<tr>
<td>Genre</td>
<td>Project defined genre (from help system)</td>
<td>Speak</td>
</tr>
<tr>
<td>Related objects</td>
<td>Webpage, podcast, photo etc.</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>Project defined subject (from help system)</td>
<td></td>
</tr>
<tr>
<td>Tag</td>
<td>Project defined subject (uncontrolled)</td>
<td></td>
</tr>
<tr>
<td>Object start</td>
<td>Time for start of part of program</td>
<td>19:00:00</td>
</tr>
<tr>
<td>Object end</td>
<td>Time for end of part of program</td>
<td>19:05:30</td>
</tr>
<tr>
<td>Annotation</td>
<td>Project annotation to entire show or part of show</td>
<td></td>
</tr>
</tbody>
</table>

Controlled keywords and genre information could originate from an existing support system within CHAOS or each project could define their own authority lists.

**Administrative Metadata**

An important part of a working metadata system is administrative metadata. In CHAOS, administrative metadata is related to each of the three metadata levels and provides information about the metadata record, i.e., when and by whom the record was created. This data can be generated automatically, e.g., from user login. Below are some examples of administrative metadata related to the archival level:
Administrative metadata at LARM and project-specific levels should also include information about when and by whom records have been edited. Here, administrative metadata does not only have to be related to the record as a whole, but can also apply to actions taken on individual metadata elements in the metadata records, i.e., who changed the title information and when.

### Annotation Support

Our proposed metadata architecture (see Figure 1) provides for different help systems to support users in their data entry and annotation process. Such support systems could take the form of standardized vocabularies or taxonomies, authority lists of person names, etc. Their main purpose is to secure uniform entry points to help users in the retrieval of radio material.

We propose two types of support systems: a centralized support system as well as project-specific support. The centralized support system allows for the importing of existing taxonomies and authority lists, such as EBU Escort (EBU, 2007) or DR index terms. In addition, each project should be able to add its own project-specific support system. Every support system should be documented by means of a coding manual. Documentation about existing standardized system, such as taxonomies, should be in the form of a reference to an existing formal documentation or standard. Project-specific support systems should provide the required documentation themselves, including information about their intended use, coding, etc. We propose that a specific interface be implemented in CHAOS for the creation and editing of project-specific support systems and the automatic importing of existing classification schemes. The "EBU Reference Data & Classification Schemes" (EBU, 2011) could be used as reference scheme for this.

### Conclusion & Future Work

The proposed CHAOS metadata scheme has been implemented in the latest release of the CHAOS infrastructure. We aim to conduct a thorough evaluation of the metadata scheme to qualify its relevance. This will include both an evaluation of the suggested metadata elements, but also of the CHAOS user interface.

### References


ICT Practices by Voluntary Groups: A Multi-Setting Study

Ali Eshraghi
The University of Edinburgh Business School
s1057376@sms.ed.ac.uk

Abstract

This poster describes my interdisciplinary doctoral research on information and communication technologies (ICTs) and voluntary sector organizations (VSOs). Despite the fact that ICTs proliferate within these organizations, the literature on the relationships between such technologies and voluntary contexts is scant in comparison to private and public domains. In particular, it is not clear how the specific characteristics of volunteers might affect the take-up and use of ICTs. Using the language and concerns of sociology of technology and by focusing on socio-technical dynamics of everyday ICTs practices, my research seeks to explore and conceptualize the ways such practices are constituted, maintained and reproduced by multiple actors and settings. The emerging themes of the research show that high turnover of volunteers and their different levels of engagement in VSOs shape the ICT-related practices. This poster also identifies the contribution of the study.

Keywords: voluntary sector organizations, information and communication technologies, organising practices, social actors, distributed settings

Introduction and Literature Review

Voluntary Sector Organizations (VSOs) serve different roles in society through self-governing mechanisms and also voluntary activities (Kendall, 2003). VSOs, like other business firms and public sector organizations, are also welcoming ICT innovations to respond to the increasing challenges of the sector such as competition for funding and volunteers [Burt & Taylor, 2000]. This increases the proliferation of ICTs in the sector and hence there is a growing consensus that we need to study ICTs and VSOs [Pereira & Cullen, 2009], in particular because such technologies alter the ways people organize their work activities [Orlikowski & Barley, 2001].

Scholars, mostly with a background in VSO fields, have provided valuable insights into the reasons of VSOs failure in reaping the benefits of ICTs (e.g. Hackler & Saxton, 2007)]. In addition, their studies have suggested some of organizational consequences for implementing such innovations (e.g. Burt & Taylor, 2003). However, there are still conceptual and methodological inadequacies within the current literature. The challenges associated with the ‘Deterministic View on ICTs’ and ‘Single Setting Research Design’ are two less-considered areas. This provides space for further studies in this area like the proposed research.

Many ICT-related VSO studies seek to explore and to theorize the impacts of technological innovations on organizations as well as to determine factors which influence processes of technology adoption and use in VSOs. Such deterministic views have not considered the co-evolution of technology and organization and hence the possibility of dual effects of technology on organizational activities is not well-understood. (e.g. Hart, 2002; Zhang & Gutierrez, 2007).

In terms of research design, the unit of analysis for studies on ICTs in VSOs has varied from individual level of analysis [Zhang & Gutierrez, 2007] to organizational [Hackler & Saxton, 2007] and to national level [Malina & Ball, 2005]. However, there is a lack of multi-level analyses that could investigate the dynamics and perhaps the mutual impacts of different actors on each other across and within VSOs. Moreover, a number of these research inquiries are also ill-contextualized.

Copyright is held by the author.
These inquiries have attempted to re-use other technology studies knowledge in VSO domain by conducting statistically-informed surveys which explore the demographics of technology adoption and use (c.f. Iverson & Burkart, 2007).

Therefore, this research aimed to explore ICT-related practices within the specific context of VSOs through an interdisciplinary research, including Science and Technology Studies (contexts), Organization Studies (practices), Computer-Supported Collaboration Work (interactions), Information Systems Research (artifacts), and more importantly Social Informatics (dynamics). The broad research question that this study tackles is: “How everyday ICT practices are constructed, maintained and reshaped in smaller voluntary organizations?”

This study narrows the scope of VSOs to those ones that rely heavily on voluntarism, though other VSOs may be better explained through other characteristics such as being non-profit making/distributing. Considering the short-term nature of doctoral studies, this research focuses on exploring the case of small scale VSOs. This gives the researcher an opportunity to build a richer picture of the object of study which is “everyday ICT-related practices” by voluntary groups. This means that this research offers a basis to find out the ways through which the voluntary nature of work affects the take-up and then the use of ICTs in organizing practices by volunteers.

Research Methodology and Theoretical Approach

As discussed earlier, the current literature on VSOs and ICTs leaves the technology as a black-box. By formulating a ‘how’ question, this study seeks to explore what happens inside the ‘black-box’. In doing so, initially, an in-depth ethnographic case study of a leading Scottish swimming club was designed to shed light on the use of ICTs by volunteers, in particular by applying the insight of the mainstream CSCW studies and IS research. However, during the earlier stages of the fieldwork, a fundamental emergent issue has arisen: people of the case study were pointing to other actors and settings beyond the case when they were being asked about their everyday work practices. This was also supported by further observations during their work with a range of ICTs. For instance, a new sports software vendor has come into the market and some people in the club were under-exploiting the functionalities of the current software with a hope to the procurement of the new system.

This issue might not be a matter of single case study; rather it seemed that its roots were in focusing on volunteers’ use of ICTs within a single setting, which was the club setting. The emergent issue, i.e., the flags made by volunteers, has encouraged me to start conversations with more diverse communities to establish an appropriated theoretical basis. This provides more insights into the research focus through different concepts such sociomaterial practices [Suchman, 2007], affordances [Bloomfield et al., 2010], entanglement [Barad 2003]. Among these, the Social Actor Model (Lamb and Kling 2003) provides a useful insight with regard to the nature of research problem. Through this model, Lamb and Kling argue that atomistic views of technology users are problematic and instead they call for a re-conceptualization of how the user is understood. They suggest that an organizational member is not just a user; he/she is a Social Actor whose interactions are enabled and perhaps constrained by a number of other external and internal entities. Using a socio-technical approach, this model seeks to explore and explain how these entities alter the use of ICTs within organizational contexts. In doing so, they characterize four ICT-related dimensions of a social actor: affiliations, environments, interactions and identities.

Using their broad and open-ended dimensions and considering the emerging voice on the significance of materiality of technological artifacts, this research seeks to conceptualize volunteers (and other VSOs members) as social actors which their practices of ICTs are not just shaped by a specific kind of technology within a limited workplace. Instead, to study such practices, we need to go beyond the boundaries of the case and also to take into account other human or non-human actors. Hence, it seems that the previous ‘flat ethnographic’ case study [Williams & Pollock, 2009] should be completed with other techniques to generate data and make sense of more distributed contexts [Monteiro et al. 2012].

As a result, this research has three pathways: first, the idea to study volunteers as users of technology in a limited setting is problematic; second, it has been argued that users are social actors that their ICTs practices are shaped by diverse internal and external entities; third, a multi-setting study is required to capture all significant key players and contexts which affect the patterns of ICTs practices. This research therefore studies both the club level and other distributed settings and actors which they are being emerged through snowballing techniques.
Preliminary Findings and Future Contributions

Within six months of ethnographic fieldwork at the club level, two central themes have been emerged. For the first theme, the generated data suggests that since the club’s volunteers are different, those who have more time to offer may also have more power to shape the general pattern of ICTs use. In particular, if a key actor decides not to use a collaborative technology, that technological system would probably fail. The second theme is based on the issues and challenges surrounding the high turnover of volunteers in smaller scale VSOs. This ‘coming-and-going’ has led to many of workarounds as people have more freedom to choose from portfolio of available technologies.

The fieldwork has also revealed other distributed settings, people and non-humans which have an influence on the club’s ICTs practices. This includes a national governing body for swimming, an online intermediary which coordinates and regulates all Scottish swimming competitions, independent practitioners and two software vendors. The future plan for the study is to continue with the ethnographic data collection from the club (shadowing, observations and talks) as well as to generate further insights into the distributed actors/settings (official reports, interviews and online materials).

This research is expected to contribute in three main domains. Firstly, consistent with the calls for multi-locale technology studies (c.f. Koch, 2007; Pollock & Williams, 2010), this research generates insights into the role of distributed contexts of ICTs practices. Secondly, although this research is informed theoretically by the social actor model, it extends the model through the application and modifications needed for the context of small VSOs (c.f. [Lamb 2005]). Finally, the study is expected to support policy-makers and practitioners in (re)defining their ICTs plans/programs.

References


Gendered Artifacts and User Agency

Andrea Marshall
Drexel University
am3432@drexel.edu

Jennifer Rode
Drexel University
jar394@drexel.edu

Abstract

In this poster we investigate the tension between feminist design ideals and issues of usability with regard to smart closets. Technological artifacts are inscribed with cultural attitudes toward gender, and in turn they can be used to reinforce societal conceptions of gender to users. We propose that smart closets, such as ClosetClique, our user prototype interface, present an opportunity to recognize gender as an integral part of design, in order to further realize feminist design goals and to more fully examine female user experience(s). Additionally, our discussion of usability offers an explicit way to explore how female users react to smart closets as gendered technological artifacts, thereby revealing that everyday contexts of use explicitly contribute to the creation of feminine sociotechnical identities.

Keywords: gender, feminist HCI, usability, participatory design

Introduction

Ubiquitous computing is increasingly capturing the attention of popular culture, as well as designers and researchers. Smart closets are technological artifacts that reflect this scholarly interest. Smart closets are technological systems with user interfaces that allow users to manage their clothing selection and perform wardrobe maintenance (Marshall and Rode 2012). Additionally, current scholarly literature within ubiquitous computing has discussed smart closets as recommender systems that can assist users with sartorial choices (Rode et al. 2012). We propose that our prototype user interface ClosetClique can expand the sartorial choices of female users. ClosetClique possesses functionalities that support female users’ agency with regard to their wardrobe selections. Our prior ethnographic research with regard to smart closets has suggested that engaging female users in the process of participatory design supports feminist design goals (Rode 2011; Rode et al. 2012). However, our research has revealed a design challenge: We examine how we can create technological artifacts that both incorporate feminist ideals and support functionalities that are influenced by concerns of usability.

There are multiple ways gender is embedded into the design of technological artifacts, which in turn reflects the social realities of various gendered identities. Harding discusses three understandings of gender: individual gender (the gender identity of a single person), gender structure (the division of labor between genders) and gender symbolism (the association of gender with characteristics other than the individual or the division of labor (Harding 1986). Rode (2011) further proposes that there is an explicit need for sociotechnical theoretical work to be performed within gender and HCI. While we each have our own gender identity, gender structure and gender symbolism act in concert as normative influences on the way we understand gender in daily life. Furthermore, the categories of masculine and feminine are socially and culturally constructed (Butler 2006; Calasibetta and Tortora, 2003; Rode et al. 2012). While each individual has her own gendered identity, socially endorsed gender categories affect how people perceive themselves as gendered bodies.

Gender is often considered to be an invisible element of design since technology is often traditionally construed as masculine (Rode 2010; Rode 2011); this viewpoint has the potential to alienate female users. One way to counteract this is through the careful scrutiny of feminine social practices embedded within smart closet technology (Marshall & Rode 2012); this requires an acknowledgement

Acknowledgements: The authors extend grateful thanks for Google and Nokia for funding our research.
Copyright is held by the authors.
that social values and cultural practices are inscribed into the designs of technological infrastructures (Star 1999). Understanding gender symbolism is a critical component of understanding how technological artifacts contribute to construction of gendered sociotechnical identities (Berg & Lie 1995; Rode 2005; Rode 2010; Rode et al. 2012). Furthermore, Berg and Lie (1995) contend that “artifacts do have gender and gender politics in the sense that they are designed and used in gendered contexts”. This supports Harding’s (1986) assertion regarding symbolic gender. Technological artifacts are situated in gendered environments. The assumptions present within these environments contribute to the understanding of masculinity and femininity through the gendered assumptions both incorporated into the design of the infrastructure and the contexts of use (Berg & Lie 1995; Marshall & Rode 2012; Rode 2011; Rode et al. 2012). Gender is culturally inscribed onto technological artifacts and reflects the values of the designers as well as the users. Since technological artifacts are “social constructs, artifacts are reservoirs of information on socio-cultural patterns but also on possibilities for change within these patterns” (Berg & Lie 1995). User experiences ascribe gender traits onto technological artifacts and reify the gendered assumptions ingrained into the design of these artifacts. Participatory design is one way to ensure that feminist goals could potentially have realizable outcomes for the building of technological artifacts such as the smart closet.

**Sartorial Choices and Symbolic Gender**

We have observed in our prior ethnographic research that women’s clothing selections are extremely personal (Rode et al. 2012); therefore, participatory design is one method of pursuing feminist goals in the construction of technological artifacts such as smart closets. Participatory design is intensely collaborative (Bardzell 2010; Rode 2011; Rode et al. 2012), which can also address concerns of usability with regard to smart closets. We suggest that needs validation testing as a participatory design method (Rode et al. 2012) is especially crucial to understanding how the functionalities of the smart closet interact with feminine contexts of use (Marshall & Rode 2012). These contexts of use in turn engage directly with feminist design goals that seek to draw attention to gender as a visible component within technological artifacts and systems. Furthermore, because these contexts of use have been disclosed via our prior ethnographic research (Marshall & Rode 2012; Rode 2011), they identify the smart closet as a site of technical negotiation between gender identity (users) and gender functionality (designers). This in turn reveals the smart closet as an active agent that assists female users in constructing their gendered selves, rather than a passive artifact or a dictatorial recommender system, the latter of which users actually dislike (Rode 2011). This collaborative process realizes feminist ideals for design whilst supporting feminine values through functionalities, such as sharing with friends and crowdsourcing through social networks connected to the prototype user interface (Marshall and Rode 2012; Rode 2011).

The symbolic gendering of artifacts occurs in conjunction with user experiences that enact gendered contexts of use (Poon 2009; Robertson 2000; Rode 2011; Rode et al. 2012; Star 1999). Our prior work suggests that young women users of our ClosetClique interface constructed gendered sociotechnical identities within familiar contexts of use (Rode et al. 2012; Star 1999). The young women users we researched in our speed dating study used our smart closet as a mediator between their sartorial choices and the construction of their gendered sociotechnical identities (Marshall & Rode 2012; Poon 2009; Rode et al. 2012). Our user scenarios emphasize that the smart closet is fully integrated into the lived realities of female users; the ClosetClique user interface functions as a nonhuman social actor that translates the private sartorial decisions of users into the performed gender identities in the public sphere (Butler 2006; Poon 2009; Rode et al. 2012).

Our fourth ClosetClique user scenario displayed below (Rode et al. 2012) shows the user Lena engaging in a sartorial decision; she is attempting to construct a gendered self for the workplace. Her smart closet becomes an active agent in this process; it utilizes visual and auditory feedback of prior occasions she has worn the dress. This triggers pleasant memories for Lena and her confidence is directly linked to this clothing choice. This particular scenario also showcases an interaction between the smart closet and the female user that results in the construction of a gendered self; this transition from the privacy of Lena’s bedroom to the public sphere of her workplace emphasizes the smart closet’s relevance as a nonhuman actor in this process.

1000
Gendered Artifacts and Identity Construction

As designers, we can attempt to socially construct the symbolic gender of technological artifacts; however, we must be mindful of binary gender norms while doing so (Butler 2006). Judith Butler (2006) argues that gender roles are treated in a binary manner and that sexualities are linked to physical bodies; these strict interpretations are insufficient for the complex gender identities that actually exist in lived realities. Technological artifacts can contribute to the social constructions of gender, and may disrupt or damage a user when they must perform articulation work regarding her gender (Robertson 2000). Technology designers are faced with the difficult task of addressing the way their designs interact with their users in light of the interplay between gender and technology. Since the social construction of gender is an element embedded into the design of technological artifacts, participatory design is one method by which feminist ideals might be realized and binary gender roles could be defied, in favor of users’ own self perceptions of gender.

Prior work on smart closets suggest that clothing practices are integrated into the social structures of nurturing and support, which are traditionally feminine values (Chia 2009; Dutcher 2009; Picken 1957; Rode 2010; Star 1999). Additionally, technological artifacts can influence the individual gender of users, but might also be symbolically gendered themselves. When we develop these artifacts, we have a responsibility to ensure that we avoid instantiating static gender roles or identities into our designs (Mainwaring, Chang & Anderson 2004). We need to fully recognize that users can fully reveal areas we might pay attention to in these circumstances. We also need to acknowledge that when users perform extended articulation work in order to accommodate smart closets that dictate gendered sartorial choices (Rode et al. 2012) the results are recommender systems that do not function as feminist technological artifacts.

We must also consider the tension between the support of usability and the reinforcement of conventional gender roles. Since feminine values can run the risk of reflecting conventional social attitudes, discussions about usability can disclose possible solutions to these essentialist views of gender (Rode 2011). We have explored this concept in our prior ethnographic work and through our user scenarios (Rode et al 2012). Furthermore, binary gender is needed to begin categorical examination of why women might be alienated from technologies conventionally perceived to be masculine (Rode 2011). It is important to note how the smart closet is an opportunity for critical design, and as designers we utilize gender as an element of design, but with the understanding that user experiences reflect the societal attitudes embedded within the infrastructures of technological artifacts (Rode 2011; Rode et al. 2012; Star 1999).

Implications for Future Work

It is a challenge of critical design that we empower users and avoid implementing our own values in our designs. We hope to implement participatory design as a method of empowerment by speaking with potential end users as we construct the smart closet, as we have attempted to apply in our prior research (Rode et al., 2012). The act of design in this instance is collaborative, and usability testing also helps to realize this cooperative goal. Less articulation work might be the result of realized understanding users’ experiences that we have identified in our prior work (Marshall & Rode 2012; Rode et al. 2012). Our goals for future work aspire to speak directly to users’ needs through the ethnographic analyses of
their lived experiences with technological artifacts. This will allow us to expand the potential of the smart closet as a feminist technological artifact.

Smart closet technologies present an exciting change to combine feminist design goals with participatory design practices. A feminist approach could help us understand more about how gender identities are influenced by technological artifacts, and give us the capacity to create technological systems that support contexts of use revealed by the users themselves through participatory design. This would allow us to support the diverse, subtle and vibrant gender identities that reflect the lived realities of smart closet users.

References

Collaborative Information Behaviour of Graduate Students in Tehran University of Medical Sciences (TUMS)

Maryam Zakerhamidi
Tehran University of Medical Sciences
Tehran, Iran
Maryamzakerhamidi@gmail.com

ABSTRACT

Collaborative information behavior is an important and growing area of research in the field of information behavior. Although collaboration is a key component of work in organizational and other settings, studies when two or more actors identify, seek, search, and use information to accomplish a task. This poster reports on a recent research investigating the collaborative information behavior of graduate students in Tehran University of Medical sciences, students of in the context of Medical science research groups projects. A total of 25 graduate students who participated in a collaborative research project were asked to complete process surveys throughout the project. Data analysis showed that collaborative information behavior dynamically occur in different stages of the project through different levels depending on the corresponding project tasks that trigger different collaborative information related activities.

Keywords: information behavior, collaborative information behavior, HCI, TUMS

Copyright is held by the author.
Sociotechnical Systems Research:
Defining, Converging, and Researching as a Community

Rebecca Reynolds
Rutgers University
rebecca.reynolds@rutgers.edu

Warren Allen
Drexel University
wsa25@drexel.edu

Ingrid Erickson
Rutgers University
ingrid.erickson@rutgers.edu

Shuyuan Mary Ho
Florida State University
smho@fsu.edu

James Howison
University of Texas-Austin
jhowison@ischool.utexas.edu

Ingbert Schmidt
Univ. Illinois Urbana-Champaign
ifloyd2@gmail.com

Abstract

The Consortium for the Science of Sociotechnical Systems (CSST) serves as a trans-discipline community, connecting like-minded scholars from many different intellectual communities. CSST brings together researchers from a wide range of disciplines to develop a common language and scholarly repertoire as we work to understand diverse sociotechnical issues. Researchers focus on improving human lives through understanding sociotechnical systems, conducting research on human activity such as collaboration, creativity, learning and economic production in domains like healthcare, education, science, leisure, and computing. This requires researchers to understand both social and technical aspects of human organization. This workshop supports continued advancement of definitions and boundaries in this area. We will engage in activities with established leaders as well as newcomers in this trans-discipline, to build understanding of factors that support the community’s cohesion, and, aim to leverage the diversity of the work being conducted by its members, to engender learning and research innovation.

Keywords: socio-technical systems research, consortium for the science of socio-technical systems, CSST

Introduction

For four years, the Consortium for the Science of Socio-Technical Systems’ (CSST) annual Summer Research Institute has offered an intensive program of exploration and discovery to educate and build community among successive cohorts of socio-technical systems researchers. Recurring workshops at the 2009, 2010, 2011, and 2012 iConferences have likewise provided a regular venue for researchers to convene and share their interests in socio-technical issues within the iSchool community.

The 2013 CSST iConference Workshop, now under the auspices of an NSF Research Coordination Network, builds on this tradition of sharing and community building by offering a full-day workshop to help both new and more seasoned socio-technical scholars to identify research opportunities, stimulate future collaborations, and contribute to the evolution of the larger socio-technical field of scholarship.

Overview

The first half of the workshop will be targeted towards newcomers as an introductory session that demystifies analytical approaches common in socio-technical work such as social shaping of technology (SST), social construction of technology (SCOT), actor network theory (ANT), social informatics,

Acknowledgement[s]: We are grateful for the support of the National Science Foundation’s Directorate for Computer & Information Science & Engineering. Reynolds, R., Allen, W., Erickson, I., Ho, S. M., Howison, J., & Schmidt, I. (2013). Sociotechnical systems research: Defining, converging, and researching as a community. iConference 2013 Proceedings (pp. 1004-1006). doi:10.9776/13186

Copyright is held by the authors.
sociomateriality, and design. The second half of the workshop will involve hands-on group activities centered around identifying research gaps and potential funding streams, applying research methods, building common resources, and determining protocols for continued community building.

The workshop will provide a unique opportunity for sharing problems and results amongst a growing research community. The workshop will help to identify opportunities, stimulate future collaborations, and contribute to the evolution of sociotechnical research. Three goals motivate the proposal for this workshop:

1) To inspire sociotechnical researchers with stories of success in outcomes and impact.
2) To identify new areas of opportunity for sociotechnical researchers.
3) To promote the sociotechnical field of research to newcomers, by reviewing its origins, summarizing its successes as a community, and communicating the current state of the field, which offers greater clarity on its definition and conceptual boundaries of this approach.

Notes, speaker presentations, and possibly video will be disseminated on the workshop’s website in order to reach outside of the group of workshop attendees. With the current funding climate and the increasingly competitive publication sphere, questions of impact are pervading the sociotechnical discipline. With this conference we aim to identify previous successes and opportunities for future successes in order to provide a firm foundation for researchers to continue in this realm of work. As sociotechnical researchers are most closely aligned with the iSchools (as demonstrated by our large workshop attendance in the last several years) we would like to spur the next year of research and grant writing at the iConference. All participants are asked to prepare a 1-page executive summary brief of their research agenda, and bring 15 copies.

### Draft Agenda

<table>
<thead>
<tr>
<th>Duration</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30-9:00</td>
<td>Informal self-introductions over coffee and pastries</td>
</tr>
<tr>
<td>9:00-10:30</td>
<td>Individual Speaker Session: Case studies of exemplary work in socio-technical systems research</td>
</tr>
<tr>
<td></td>
<td>• 9:00-9:15, Brief introduction to agenda, co-organizers</td>
</tr>
<tr>
<td></td>
<td>3 presentations:</td>
</tr>
<tr>
<td></td>
<td>• 9:15-9:45 Speaker 1, Bonnie Nardi, Professor in the Department of Informatics in the Donald Bren School of Information and Computer Sciences at the University of California, Irvine (confirmed)</td>
</tr>
<tr>
<td></td>
<td>• 9:45-10:15 David Ribes, Assistant Professor in the Communication, Culture &amp; Technology Program at Georgetown University (confirmed)</td>
</tr>
<tr>
<td></td>
<td>• 10:15-10:45 Third speaker TBD</td>
</tr>
<tr>
<td>10:45-11:00</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00-12:15</td>
<td>Master Class, linking features of the case examples to the &quot;socio-technical system&quot; domain of scholarship:</td>
</tr>
<tr>
<td></td>
<td>• Steve Sawyer, Associate Dean for Research, Professor, Syracuse University School of Information Studies to facilitate</td>
</tr>
<tr>
<td></td>
<td>• 11-11:45 Moderated panel discussion among the earlier speakers drawing out the sociotechnical themes in their work.</td>
</tr>
<tr>
<td></td>
<td>o Includes small group rubric activity: Identify the socio-technical elements of the earlier research exemplars</td>
</tr>
<tr>
<td></td>
<td>• 11:45-12:15 Open Q &amp; A and dialogue</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>12:15-1:30</td>
<td>Lunch with Groups</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1:30-3:00</td>
<td>Mapping domains and terrains within pre-determined groups</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>3:00-3:15</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>3:15-4:45</td>
<td>Continuation, Mapping domains and terrains within pre-determined groups</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>4:45-5:00/5:15</td>
<td>Final wrap-up and sharing time.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Educating E-Discovery Litigation Support Professionals

Jason R. Baron  
University of Maryland  
jason.baron@nara.gov

Douglas W. Oard  
University of Maryland  
oard@umd.edu

Abstract

E-Discovery is the process in civil litigation through which one party to a lawsuit requests digital evidence from the other. In this workshop we will learn about e-discovery education offerings in a variety of academic settings, and we will discuss the place of education for e-discovery professionals in iSchools.

Keywords: e-discovery, education

Introduction

The past decade has seen explosive growth in the need for “discovery” of digital information held by one party in to lawsuit in response to a request posed by a counterparty. In less than a decade, “e-discovery” has become a multi-billion dollar industry and a substantial driver in the cost of civil litigation. No current degree programs at any level are designed specifically to educate e-discovery professionals, so employers presently recruit from many source disciplines (e.g., forensics, business schools, and law schools). iSchools seem like natural homes for such programs: some of us offer joint degrees with our law schools, others share a campus with one of the 42 law schools that offer an e-discovery course, and still others have faculty who have published on e-discovery.

Overview of the Program

The program will include presentations from e-discovery thought leaders, from professionals who employ litigation support staff, and from faculty who have taught e-discovery courses. Small-group sessions focused on curriculum design for one-course offerings, and on program design for more extensive offerings, will offer opportunities for close engagement between participants and presenters.

Outcomes

Expected outcomes of this workshop will be an understanding of this opportunity, some insight into what shape such programs might take, and a community of like-minded faculty who can share ideas and resources.

Copyright is held by the authors.
Abstract

Digital age youth has emerged as a focus of interest to researchers across many disciplines. iSchool researchers address youth through designing applications and writing code, policymaking, understanding youth behavior and resources in myriad situations—including school, family, and neighborhoods. Despite the common interest in youth, iSchools have yet to come together as a field and move forward around a shared agenda. The goals of the 2013 workshop are to bring together researchers who are interested in digital youth and interested in being at the outset of a new digital youth community. The workshop will consist of one-minute “lightning” presentations followed by an unConference format where participants come together around working groups of interest. The 2013 iSchool workshop is the first in a three-event effort to establish an iSchool voice as leaders on the international stage around promoting digital youth. The first event, iConference 2013, is aimed at drawing a diverse initial community that will further comprise a steering group to help plan the second event, a Digital Youth Summit 2014 (in Seattle) that will be sponsored and attract participants from industry, government, nonprofits, education, and media. The third event will return to the iConference 2015 for reporting out.

Keywords: digital, youth, researchers, designers, policy
An Introduction to Policy Informatics

Erik Johnston
Arizona State University
Erik.Johnston@asu.edu

Kevin DeSouza
Arizona State University
Kevin.Desouza@asu.edu

Tanya M. Kelley
Arizona State University
tanya.m.kelley@asu.edu

Abstract

In a complex world generating copious amounts of knowledge, new perspectives are needed for individuals and publics to understand, deliberate, and organize collectively to build public goods. This workshop provides an interactive introduction to the theory, methods, and actionable applications of policy informatics. Policy informatics is the study of how advances in computation, information, and communication technologies are leveraged to understand and address complex policy and administrative problems and realize innovations in governance processes and institutions. Participants will engage with content and other participants. Four themes are covered: 1) theories useful to understanding policy informatics, 2) information, tools, and changing norms in the public sector, 3) interactive simulations and models, 4) participatory platforms – design, governance, and outcomes. There will be a presentation - interaction format using cutting edge examples and activities, and participants will learn techniques and methods for solving problems in the context of the values and objectives of government.

Keywords: policy informatics, governance, research methods

In a complex world that generates copious amounts of knowledge, new perspectives are needed for individuals and publics to understand, deliberate, and organize collectively to build public goods. One reason government is focused on information management and analysis is that the interactions norms present in open information environments, such as crowdsourcing and collective intelligence, are aligned with civic values present in public service. Another reason is that people are becoming used to interacting with both public and private organizations through technological platforms. For government to increase its legitimacy and sustainability there must be an understanding of how people interact in a technology mediated world and governments must leverage these advances to develop strong relationships with the public. This half day workshop will provide participants an interactive introduction to the theory, methods, and actionable applications of the emerging field of policy informatics. Policy informatics is the study of how advances in computation, information, and communication technologies are leveraged to understand and address complex policy and administrative problems and realize innovations in governance processes and institutions. It is built on the fundamental premise that information can be efficiently and effectively mobilized to enable evidence-driven policy design, implementation, and analysis. Policy informatics advances the goal of building public institutions that are transparent, collaborative, and participatory.

Policy informatics is inherently interactive and the workshop participants will have the opportunity to engage with each other and the material throughout. The workshop will cover four themes: 1) theories useful to understanding policy informatics, 2) information, tools, and changing norms in the public sector, 3) interactive simulations and models, and 4) participatory platforms – design, governance, and outcomes. There will be a presentation - interaction format for each theme, using cutting edge examples and activities occurring at all levels of government.

Copyright is held by the authors.
• **Theories useful to policy informatics**

In this section we will show how policy informatics builds on theories from multiple disciplines to form a common set of principles. These theories including decision theory, rational choice, game theory, public administration, data mining, modeling and simulation, model validation, probability, behavioral economics, organizational behavior, complex adaptive systems, information visualization, human computer interaction, and more.

• **Information, tools, and changing norms**

There is a great deal of publicly available information available that should be considered a natural resource, and if left dormant, a wasted government resource. There will be a discussion of the current state of online government and identification of information repositories. There are tools available to process information apart from traditional methods of analysis. Participants will learn about data processing tools, such as visualizations, that provide researchers, practitioners, and members of the public new understanding of the dynamics of an issue. There will be a discussion about how access to open information has been changing government, research, and public norms of interaction.

• **Interactive simulations**

Policy informatics incorporates more than accessing and processing publicly available data. Simulations can give researchers and participants new insights into a problem or situation that can inform current and future actions. An agent-based model about building collaborations will be presented to illustrate how simulations complement more traditional research approaches. Participants will learn how interacting and building models changes how systems are articulated. Participants will have a structured interaction with a model to deliberate on water policy in arid climates under uncertain conditions.

• **Participatory platforms- design, governance and outcomes**

The public sector is particularly well suited to creating and supporting interactions that can contribute to the overall welfare of the public. Participatory platforms like Challenge.gov, 10,000 Solutions, and our work designing the White House Policy Challenge are useful examples for creating new avenues for direct and productive interactions. We explore the emergence of participatory platforms and highlight the importance of research for informing decisions about motivations, design choices, and building communities of participants that can be used to create meaningful participatory interactions between government and the public.

This workshop is relevant to the iConference by outlining how the policy informatics is producing actionable research to increase understanding, improve effectiveness, and build a collaborative community of research and practice. Participants will learn a variety of methods and techniques used for processing information and solving problems in the context of the values and objectives of government. It also demonstrates how informatics technologies can lead to meaningful, interactive, multi-directional relationships across disciplines. The workshop should target around 30-40 participants.
Big Social Data for Social and Information Scientists

Sean Goggins
Drexel University
sgoggins@drexel.edu

Christopher Mascaro
Drexel University
cmascaro@gmail.com

Nora McDonald
Drexel University
norakmcdonald@gmail.com

Alan Black
Drexel University
aeblack@gmail.com

Guiseppe Valetto
Drexel University
valetto@cs.drexel.edu

Contact:
icferencebigdata@gmail.com

Abstract

This workshop aims to help build Big Social Data collection, management, analysis and visualization research competencies in the iSchool Community. We will share our strategies, methods, and tools using data of interest to participants. We propose three paths to participation. First, participants interested in big data who do not yet have well formed research questions, will be provided with public data sets collected from Twitter. Second, participants may submit research questions and corresponding search terms in advance of the workshop; we will use our existing Twitter data collection infrastructure, TwitterZombie (Black, Mascaro, Gallagher, & Goggins, 2012), to collect public data on their behalf under our existing IRB at Drexel University. Third, participants may submit a structurally representative, anonymized sample of corpora to analyze using one of our published methodological approaches; we will evaluate the structure of the data to determine appropriateness.

Keywords: big data, twitter

Introduction

When people interact through web, mobile device and distributed sensors, digital traces of these interactions are left behind. These historic interactions become more easily quantifiable through digitization and sharing of document and image archives. As a consequence, we face a transformative and disruptive data deluge, from which new scientific, economic, and social value can be extracted. “Big Data” becomes “Big Social Data” when it arises as a result of human-to-human interaction, and herein lies the key to unlocking important insights about social processes operating at a worldwide scale as they unfold over time. The potential to understand social processes through electronic trace data is greater now than it has ever been in the history of mankind because of the ubiquity of social media. Interactions can be of many kinds (conversation, exchange, response, relationship) and observed at the individual, group, organizational, and national level.

Data Collection and Management

During the morning of the workshop we will start by giving an overview of the Twitter API architecture, how our collection tool works (how it compares with other tools that collect from Twitter API), key stumbling blocks and related analytic solutions. No advanced technological knowledge is required. We will use this basis for the late morning and afternoon sessions where we dig into data.
Collaborative Analysis and Visualization

Following this brief introduction to data collection and management strategies for Twitter, we will get deep into collaborative analysis. We will share our analysis infrastructure and help researchers better understand how to approach answering their research questions with either the data we collected for them with TwitterZombie, or using their own data, collected in another manner.

Participant defined datasets will be provided during the workshop; participants will learn how to collect, manage, analyze and visualize their own data using R and Gephi. Participants will also be provided overviews of each other’s data. During this time we anticipate that many researchers will ask questions we would have not have considered and this collaboration will greatly enhance the area of collaborative data science.

Workshop Format and Schedule

We aim to mix a series of short lectures with discussion and hands on activity to capitalize on the unique experiences, expertise and interest of iConference participants.

Hour 1: Personal Introductions and Orientation to the day:
Individuals with similar research interests will be identified prior to the workshop. Introductions will focus on individual descriptions of their interests, experiences and goals for the workshop.

Hour 2: Collection and Analytical Infrastructure Overview:
We will highlight how our infrastructure is different (both better and not as robust at times). This allows the researchers to report their findings and analysis with a high specificity. We also want individuals to understand some of the technology and how easy it can be to setup our infrastructure in the cloud. Participants will be introduced to approaches they can use to analyze data collected using our data collection and management system.

Hour 3 – Hour 6:
We aim for a collaborative environment where individuals will work in groups to answer questions related to their own research questions, using data relevant to them. This is the central activity in the workshop, and will be conducted in an active, hands on manner.

Hour 7: Group Presentations:
As we have spent a few hours in data we find it beneficial that individuals or groups should report back what they found and future directions.

Goals and Conclusions

The workshop structure is designed to foster significant discourse among participants that we hope will lead to long-lasting collaborations. We also provide data of interest to the participants and tools to do so. This allows for the possibility of significant production of scholarship as a result of their participation. Our intended goals for this workshop are to create a working community interested in similar technological problems that instead of being focused on technical solutions are focused on asking social science research questions and having the capability to answer them.

Reference

Data Management Planning:
Where We’ve Been and Where We Are Going Next

Carly Strasser
California Digital Library
carly.strasser@ucop.edu

Martin Donnelly
martin.donnelly@ed.ac.uk

Sherry Lake
sah@virginia.edu

Abstract

This workshop provides overview, analysis and opportunity for discussion of recent developments in data management planning policies and tools in the UK, US, and other countries. Organizers and speakers include representatives from DMP Tool (USA) and DMP Online (UK). We will first update attendees on current data management policies. We will then describe and discuss the current landscape of data management planning, including all relevant stakeholders (funders, institutions, publishers, and repositories). The second half of the workshop will be a facilitated discussion of the future of data management, with an emphasis on what services and materials will better enable good data stewardship practices among researchers.

Keywords: data management, research data, services, software tools

Proposed Format

1. Welcome and workshop overview (organizers)

2. Policies update
   a. Short sessions including updates on funder policies from UK, US, Europe & Australia
   b. Issues covered will include:
      i. UK policy changes
      ii. Changes to US funding agencies
      iii. Publisher policies, including the new Journal Research Data Policy Bank (JISC), Dryad’s Joint Data Archiving Policy, and others
   c. Reiteration of the overall goals of research data management

3. Tools update: DMP Online and DMP Tool
   a. Short demonstrations of both tools
   b. DMPTool updates will include
      i. plans for new functionality
      ii. plans for funding
      iii. steps for governance and broader community engagement
      iv. steps for education and outreach to libraries
   c. DMP Online updates will include
      i. descriptions of future plans
      ii. updates on liaison with funders, outreach, education
      iii. Discussion of institutional policies and DMP-related work
4. Roundtable with workshop participants
   a. What training is needed to support data management and data policies.
   b. What resources are already available.
   c. How well have the existing policies worked?
   d. What changes should be made for the future?
   e. What help do people need in meeting data-related requirements?
   f. How should such support be funded? (Ideal world versus real world)
   g. What are the best mechanisms for working with research funders to influence their policy developments, and make them known to the research community?

Goals or Outcomes

The goals of this workshop are to:

1. Explore the differences and similarities in data policies between countries
2. Demonstrate the use of two tools for data management
3. Discuss the challenges of and barriers to data management for researchers, and understand how best information professionals can address these barriers
4. Provide attendees with the opportunity to discuss the current data management landscape
5. Allow attendees to test out the available data management tools (DMPTool, DMP Online)

Relevance to the Conference

The workshop will appeal to the iConference participants as it deals directly with policies and tools need to support data collection and management in the research community. The participants will see and hear how international partnerships were instrumental in the development of the DMPTool. Not only is the topic a cutting-edge issue for information technology & research, it will also be of immediate use to the target audience, who are very likely to be users of DMP support tools. The discussion format will allow for audience participation, giving those attending a forum to share ideas on how best to approach data management support and services in the 21st century.

Half or Full day

Half day

Expected/preferred number of participants

Maximum of 50.
Filling the Workforce Gap in Data Science and Data Analytics

Catherine Blake  
Graduate School of Information and Library Science (GSLIS)  
Center for Informatics Research in Science and Scholarship (CIRSS)  
University of Illinois  
clblake@illinois.edu

Jeffrey M. Stanton  
School of Information Studies  
Syracuse University  
jmstanto@syr.edu

AnnaLee Saxenian  
School of Information and Department of City and Regional Planning  
UC Berkeley  
anno@ischool.berkeley.edu

Abstract

The goal of this workshop is to provide a forum for iSchool faculty who are developing programs in data analytics, eScience, eResearch, big data, and cyberinfrastructure to discuss best practices with respect to preparing students to fill the workforce need for managers and analysts to analyze big data and make decisions based on their findings.

Keywords: Data Science, Big Data, Socio-technical data analytics, Information and Data Science

Changes in how information is created, disseminated and re-used provide new opportunities to intervene within the information lifecycle to ensure that the information created today will be available for centuries to come. Jim Gray, the first advocate of data intensive science, urges us to “support the whole research cycle – from data capture and data curation to data analysis and data visualization.” (Hey, Tansley, & Tolle, 2009). The “data deluge” (Hey & Trefethen, 2003) is now a fundamental characteristic of e-science and “big science,” especially in disciplines such as in cancer (e.g., National Center for Biotechnology Information), astronomy (e.g., the Sloan Sky Survey), and atmospheric science (e.g., coupled climate models). The transition towards data-intensive science has created a critical shortage of both a knowledgeable workforce and best practices to address the challenges of data management, curation, and analysis (Kim & Stanton, 2012; Kim, Addom, & Stanton, 2011; Palmer et al, 2007; Palmer, Renear, & Cragin, 2008) and information synthesis (Blake & Pratt, 2006a, 2006b).

Programs such as “DataNet” have been created to integrate “library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise to provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline” (NSF, 2007). The Institute of Museum and Library Services (IMLS) has also invested heavily in educational initiatives that train data curation leaders around the country. Now that data curation training programs are well underway, it is time to focus our attention on workforce gaps in the latter activities in the information lifecycle such as in the McKinsey Global Institute report published earlier this year, which stated that “A shortage of the analytical and managerial talent necessary to make the most of big data is a significant and pressing challenge …”(Manyika et al., 2011, p3). The report goes on to say that “The United States alone faces a shortage of 140,000 to 190,000 people with deep analytics skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings” (Manyika et al., 2011, pg 104).

The goal of this workshop is to provide a forum for iSchool faculty who are developing programs in data analytics, eScience, eResearch, big data, and cyberinfrastructure to discuss best practices with respect to preparing students to fill the workforce need for managers and analysts to analyze big data and make decisions based on their findings. Questions of interest include (but are not limited to):

- What differentiates programs in this area (e.g. a program in big data vs. one in cyberinfrastructure? data analytics? eScience?)
- To what extent can big data be incorporated into the iSchool classroom experience?
- What opportunities might exist to share data and teaching models between institutions?
• For some students analytics is a very different way of thinking about information. What teaching strategies have been successful in bringing students up the learning curve without sacrificing intellectual rigor?
• To what extent can faculty infuse their analytics research into the classroom setting?
• iSchool programs in data science and data analytics tend emphasize the context in which data is collected and situate analytics within the broader work-flows in science, business and the community. To what extent can this balance between social and technical aspects of big data be maintained?
• What makes iSchools uniquely prepared to address research and educational requirements of data science, data analytics, cyberinfrastructure and eResearch?

The workshop organizers provide three different iSchool flavors of data science and analytics at different stages of development. Specifically, Dr. Blake will discuss the MS and PhD. Program Socio-technical Data Analytics at Illinois, Dr. Stanton will discuss the CAS Data Science Program at Syracuse and Dr. Saxenian will discuss the MS program in Information and Data Science Program at UC Berkeley. Each organizer will present their respective program and address one or more of the questions above and thus demonstrate that each iSchool brings a different perspective to the big data workforce need. With that said the iSchools emphasis on people, information and technology makes them well prepared to address both the research and educational gaps in data science and data analytics.

This workshop will be of interest for students who are interested in learning more about how training needs in big data are being addressed in iSchools, and for faculty who are developing new programs. We will ask participants to submit a two page position statement describing the research and educational issues and one or more of the questions outlined above before the workshop. After the initial organizer presentations, workshop participants will be broken into small groups to address the questions above and in particular, identify gaps between training needs and existing programs. Group leaders will then report back to the group.

References


The field of Scientometrics is concerned with the analysis of science and scientific research. As science advances, scientists around the world continue to produce large numbers of research articles, which provide the technological basis for worldwide collection, sharing, and dissemination of scientific discoveries. Research ideas are generally developed based on high quality citations. Understanding how research ideas emerge, evolve, or disappear as a topic, what is a good measure of quality of published works, what are the most promising areas of research, how authors connect and influence each other, who are the experts in a field, what works are similar, and who funds a particular research topic are some of the major foci of the rapidly emerging field of Scientometrics.

Digital libraries and other databases that store research articles have become a medium for answering such questions. Citation analysis is used to mine large publication graphs in order to extract patterns in the data (e.g., citations per article) that can help measure the quality of a journal. Scientometrics, on the other hand, is used to mine graphs that link together multiple types of entities: authors, publications, conference venues, journals, institutions, etc., in order to assess the quality of science and answer complex questions such as those listed above. Tools such as maps of science that are built from digital libraries, allow different categories of users to satisfy various needs, e.g., help researchers to easily access research results, identify relevant funding opportunities, and find collaborators. Moreover, the recent developments in data mining, machine learning, natural language processing, and information retrieval makes it possible to transform the way we analyze research publications, funded proposals, patents, etc., on a web-wide scale.

The workshop aims at bringing together researchers with diverse interdisciplinary backgrounds interested in mining large digital libraries and other relevant databases. The topics of interest include, but are not limited to:

- New approaches to measuring the impact of research publications as well as the impact of researchers in a particular field of study.
  - Identifying influential authors, experts, and collaborators within or across disciplines
  - Modeling the referencing behavior across disciplines
  - Automatic citation recommendation

Keywords: bibliometrics, digital libraries, citation analysis, mining citation graphs

Copyright is held by the authors.
Mining large digital libraries of scientific publications and linking to other databases such as funded proposals and patents:
- Identifying research trends and topics
- Extracting relevant information from research articles
- Scaling up machine learning algorithms to large datasets
- Classification and clustering of scientific publications, funded proposal, patents, etc.
- Linking together various entities, e.g., articles with articles by similarity, articles with their corresponding presentation slides, articles with the corresponding funded proposals.

Presenting open-access, novel datasets (e.g., based on Wikipedia, DBpedia, United States Census Bureau data) that can be linked to entities in digital libraries, and can help researcher develop novel technologies for analyzing scientific publications.
Help iDigBio Reveal Hidden Data:
iDigBio Augmenting OCR Working Group Needs You

Deborah Paul  
Florida State University/iDigInfo, iDigBio  
dpaul@fsu.edu

P. Bryan Heidorn  
University of Arizona/School of Information and Library Science  
heidorn@email.arizona.edu

Jason Best  
Botanical Research Institute of Texas (BRIT)/Biodiversity Informatics  
jbest@brit.org

Edward Gilbert  
Symbiota/Lichen, Bryophyte and Climate Change TCN  
egbiodiversity@gmail.com

Amanda Neill  
Botanical Research Institute of Texas (BRIT)/BRIT Herbarium  
aneill@brit.org

Gil Nelson  
Florida State University/iDigInfo, iDigBio  
gnelson@bio.fsu.edu

William Ulate  
Missouri Botanical Garden/Biodiversity Heritage Library (BHL)  
william.ulate@mobot.org

iDigBio Seeks to Improve OCR Output and Parsing of OCR Output for Faster, More Efficient, Cheaper Natural History Collections Specimen Label Digitization

Integrated Digitized Biodiversity Collections, iDigBio, is funded under the National Science Foundation's NSF Advancing Digitization of Biological Collections (ADBC) program to help hundreds of natural history museums get specimen data out of millions of drawers and off of specimen labels into an integrated database for everyone. Over 130 museums are working together, funded as Thematic Collection Networks (TCNs), to capture standardized data to send to iDigBio's HUB, Home Uniting Biocollections cyberinfrastructure. Optical Character Recognition (OCR) and OCR output analysis play an important role in many museum object-to-image-to-data digitization workflows and are integral to several of the current TCN digitization projects. The iDigBio Augmenting OCR working group (AOCR) was formed to develop a multi-faceted approach to improvement of OCR strategies, including investigation of image segmentation, autocorrection of typographical errors, semantic autocorrection, autonormalization, automated text segmentation, generating consensus records, and user interfaces.

Keywords: OCR, natural language, information analysis, machine language, iDigBio hackathon

Introduction

The iDigBio AOCR working group is actively seeking collaboration within the broader community, or those outside of natural history collections, to benefit from collective wisdom and experience in order to find ways to make digitization faster, more efficient, and effective in creating a searchable online specimen-based data-set that is fit-for-research-use. The working group will present a concise overview of the digitization workflow from natural history museum specimen label to a complete digitized data record to a linked digital database. Each panelist is presenting a distinctly different part of the digitization process and unique challenges and benefits of OCR, image analysis and processing, and natural language processing.


Copyright is held by the authors.
iDigBio AOCR Working Group Needs You

Museum specimen data comes in a variety of formats including: field notebooks, card files, paper packets and paper sheets. The data may be hand-written, typed or printed (Figure 1). The iDigBio Augmenting OCR working group is focusing its initial efforts on typed or printed text (Figure 2). Data shared in these various formats, valuable data that is not available in any other location or format, usually includes the "who, what, where, when and how" data from a collecting event that scientists need for research. From fish to fowl, flora to frogs and fossils, it is estimated that less than 10% of this specimen data in the U.S. is digitally accessible (Beach et al., 2010).

Figure 1. Cerceris compacta specimen (a wasp) and associated labels from U.C. Berkeley, Essig Museum of Entomology. Used with permission.

Figure 2. Baccharis salicina herbarium specimen label, Botanical Research Institute of Texas (BRIT), Copyright 2012 by BRIT Herbarium. Used with permission.

We will present issues and projects relating to augmenting OCR at a half-day iConference workshop on Tuesday afternoon, February 12th, 2013 (Table 2). A panel (Table 1) consisting of key members of the AOCR working group will present their digitization projects and challenges by illustrating how OCR and OCR output fit into their current workflows. Following the panel session, breakout groups will be formed to engage in further discussion around the main components of the digitization process. The breakout groups will document suggestions for new approaches to the digitization process. These might include for example new methods for image segmentation, improved OCR methods, innovative restructuring of the OCR text for database entry and other steps in the digitization process. The exact charges for the breakout groups will be adapted to fit the skills of the attendees.

An Augmenting OCR Hackathon is scheduled for Wednesday and Thursday, February 13 - 14, 2013 -- held and hosted by iDigBio and the Botanical Research Institute of Texas (BRIT) in Fort Worth, concurrent with the 2013 iConference. Programmers and others from inside and outside the museum community are welcome. On Friday, February 15th, key members of the panel will return to iConference 2013 to share the outcomes from the hackathon in a 45 minute 2013 iConference Alternative Event discussion session to report back to the iConference workshop participants.

Sharing our gathered knowledge for the benefit of all is part of the overall iDigBio mission. The iDigBio Augmenting OCR working group and other iDigBio working groups are actively working on disseminating documentation of their activities and knowledge to the wider biological collections community, and beyond. Visit the new Augmenting OCR Wiki to learn more.

To facilitate access to OCR services, iDigBio is currently developing cloud-oriented services for its users. In the context of OCR workflows, these services can include common web services hosted by iDigBio and academic or commercial partners, as well as providing users and developers with the ability to develop, configure, package, and disseminate new and experimental services by creating virtual appliances.

We are especially pleased to note the conference themes of Data, Innovation and Wisdom. These align perfectly with the overall mission of iDigBio and the TCNs to provide fit-for-research-use data.
while seeking ways, means, and people to help get that data into databases more efficiently by reaching out beyond our community's borders. The working group is excited to present their work to the iSchools members and is looking forward to finding new research partners.

Table 1
Panelists from the iDigBio Augmenting OCR Working Group

<table>
<thead>
<tr>
<th>Name</th>
<th>Title of talk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deborah Paul &amp; Gil Nelson</td>
<td>Introducing <a href="#">iDigBio</a> and the Augmenting OCR Working Group</td>
</tr>
<tr>
<td>Amanda Neill</td>
<td>Digitization of biocollections: a grand challenge in scope, scale, and significance</td>
</tr>
<tr>
<td>Jason Best</td>
<td>The Apiary Project: a workflow for text extraction and parsing for herbarium specimens</td>
</tr>
<tr>
<td>Edward Gilbert</td>
<td>Symbiota: Creating an OCR and NLP enabled user interface and workflow to efficiently digitize 2.3 million lichen and bryophyte specimens</td>
</tr>
<tr>
<td>Bryan Heidorn</td>
<td>HERBIS/LABELX: Machine Learning Approach to Parsing OCR Text</td>
</tr>
<tr>
<td>William Ulate</td>
<td>Linking Data: <a href="#">Biodiversity Heritage Library</a>: supporting knowledge discovery from digitized content</td>
</tr>
</tbody>
</table>

Table 2
Workshop Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>130 – 145</td>
<td>iDigBio and the Augmenting OCR Working Group</td>
</tr>
<tr>
<td>145 – 315</td>
<td>Panelists' Talks, each 15 – 20 minutes</td>
</tr>
<tr>
<td>315 – 500</td>
<td>Breakout Groups – Your Innovation and Wisdom</td>
</tr>
<tr>
<td>500 – 530</td>
<td>Wrap-up: Report Backs from the Breakout Groups</td>
</tr>
</tbody>
</table>

References


Acknowledgements: iDigBio is kindly funded by a grant from the National Science Foundation's Advancing Digitization of Biological Collections Program (#EF1115210). To each and every member of the Augmenting Optical Character Recognition working group, many kind thanks to you: Robert Anglin - Symbiota Developer; Jason Best - Biodiversity Informatics Director (BRIT); Renato Figueiredo - iDigBio IT; Edward Gilbert - Developer of Symbiota, Programmer; Nathan Gnanasambandam - Senior Research Scientist, Xerox; Stephen Gottschalk - Curatorial Assistant, New York Botanical Garden; Elspeth Haston - Assistant Curator: Digitisation, Royal Botanic Garden Edinburgh; P. Bryan Heidorn - Director School of Information and Library Science, University of Arizona; Daryl Lafferty - SALIX Developer; Peter Lang - Pre-Sales Representative, ABBYY USA; Gil Nelson - iDigBio Digitization; Deborah Paul - iDigBio User Services; Anna Saltmarsh - Digital Collections Manager, Kew Herbarium; Nahil Sobh - InvertNet Developer, Researcher, University of Illinois at Urbana-Champaign; William Ulate - Technical Director, Biodiversity Heritage Library; Kimberly Watson - Curatorial Assistant, New York Botanical Garden; and Qianjin Zhang - Masters Student, University of Arizona. We must also thank Peter Oboyski, Elvis Hsin-Hui Wu, Karl-Heinz Steinke, Dean Pentcheff, Matt Yoder, Hilmar Lapp, the HIP working group, Paul Morris, John Pickering, John Deck, the iDigBio staff, BRIT staff and others for their input on workshop, hackathon and alternative event inter-related content and direction as well as help with community outreach.
From iSchools to US Ignite: Strategies for Fostering Inclusive Gigabit Networks

Jon P. Gant
University of Illinois, Urbana-Champaign
jongant@illinois.edu

Abstract

The goals of this interactive, half-day workshop are to raise awareness of the U.S. Ignite initiative and the opportunities it provides to collaborate among iSchools as well as communities. U.S. Ignite brings together high-speed broadband resources to create test beds at a national scale. The initiative uses the Global Environment for Network Innovation (GENI), an NSF-funded cloud network that provides a fast virtual laboratory to develop new applications to address societal needs in healthcare delivery, advanced manufacturing, education and work force development, economic development, disaster response, and energy use. How can iSchools help libraries play a leading role in building inclusive next-generation Internet applications and services? We will showcase specific applications from communities, including Chattanooga and Cleveland, and projects such as the Video Mosaic Collaborative at Rutgers University. Municipal and institutional variations generate distinct local strategies to increase social inclusion in the digital world.

Keywords: broadband, GENI, high-speed networks, US Ignite, libraries

From iSchools to US Ignite: Strategies for Fostering Inclusive Gigabit Networks

The iSchools in the United States are at a critical juncture as the build-out of broadband infrastructure is being completed in 2013. Broadband creates opportunities and challenges for educational institutions concerned with interconnections among people, information and technology. Not as well known, but perhaps even more impactful, is the U.S. Ignite initiative, launched in June 2012. U.S. Ignite promotes “leadership in developing applications and services for ultra-fast broadband and software-defined networks” with the aim to address crucial social needs, such as health information and education. The Graduate School of Library and Information Science (GSLIS) at the University of Illinois, Urbana-Champaign, received funding for one year from the U.S. Institute for Museum and Library Services (IMLS) for “Inclusive Gigabit Libraries,” a project to examine the ways in which libraries and other community anchor institutions can provide the facilities, skills, and incentives for their users to work together to address common concerns, such as workforce development, public safety, and safe energy. Our formal grant partners are the Office for Information Technology Policy at the American Library Association, and U.S. Ignite. I am serving as the PI for the grant and will facilitate this interactive workshop. We are eager for input from iSchool members, in order to enrich our final white paper, which will be widely distributed in professional and policy circles by summer 2013.

Three brief presentations, by myself and our partners, will introduce participants to:

- Digitally inclusive communities and the role of the public library
- State of broadband access for libraries today
- Future of broadband access with the build-out of gigabit communities
- US Ignite Initiative and the opportunity for libraries
- The technology behind the fiber, in very accessible terms
- Case studies of “gig cities,” Chattanooga, TN; Cleveland, OH, and Kansas City, as well as relevant projects such as the Video Mosaic Collaborative at Rutgers University

Acknowledgement: U.S. Institute for Museum and Library Services, RE-00-12-0015-12
Copyright is held by the author.
Presentations

1:30-1:50pm  Introduction to gigabit speed networks and the IMLS grant, “Inclusive Gigabit Libraries” by Jon Gant

1:55-2:15pm  Presentation by US Ignite staffer (possibly via Skype) about US Ignite initiative, with five minutes for questions

2:20-2:40pm  Presentation by ALA Office for Information Technology Policy (Larra Clark or other OITP staffer, possibly via Skype) about libraries and high speed Internet networks

Break —

3-3:15pm  Setting up the fishbowl with guidelines and discussion topics

Fishbowl Discussion

3:15-3:45pm  How can iSchools serve as incubators for inclusive uses of high-speed networks? What are the challenges and opportunities to using gigabit speed networks for teaching and research in iSchools?

3:45-4:15pm  How can we leverage the opportunities to most effectively engage iSchools with community anchor institutions, such as libraries?

4:15-4:45  How can we mitigate the challenges of these networks to make collaborations most effective?

4:45-5:30  Wrap-up and evaluation

The goals of this workshop are to raise awareness of U.S. Ignite and the opportunities it provides to collaborate among iSchools as well as among communities. “Stitching together existing ultra-fast and software-defined networks across campuses and communities” creates exciting opportunities to incubate not only technical applications but also inclusive social structures. The iSchools community is well-positioned to contribute ideas and insights about the US Ignite platform that has been created to serve as a testbed within cities, as well as at a distance. Each institution will bring distinct local approaches to the conversation that will inform the final report and the discussions and web-based materials on which the report will be based. These institutional variations will generate strategies to increase social inclusion in the digital world.
iSchools Building on the Strengths Found in the Convergence of Librarianship, Archival, and Museum Studies to Improve the Education of Managing Digital Collections

Anna Maria Tammaro  
University of Parma  
annamaria.tammaro@unipr.it

Vittore Casarosa  
ISTI CNR Pisa  
casarosa@isti.cnr.it

Seamus Ross  
University of Toronto  
Faculty of Information  
seamus.ross@utoronto.ca

Heather Moulaison  
University of Missouri  
moulaisonhe@missouri.edu

Terry Weech  
University of Illinois Urbana Champaign  
weech@illinois.edu

Frederick Lugya  
Makerere University Library  
Uganda  
flugya@gmail.com

Abstract

The half day workshop will explore ways iSchools can participate in research and education relevant to the employment of information professionals in libraries, archives, and museums to provide access to and preservation of the collections held by these institutions. Presently, the education of information professionals for each of these institutions has primarily separate and diverse tracks and traditions. While some schools educating librarians do offer courses that include discussion of the application of library and information science skills to archival and museum collections, few offer full specializations in this area. Archival education has traditionally been based on degrees in history, but as the archival collections have utilized more digitalization to provide preservation and expand access to the content of collections, the need for staff with more knowledge of skills related to the preservation and expansion of access to archival collections has been recognized. Similar recognition of the importance of information profession skills is being recognized in museum studies, an area that in the past has depended on specialized subject discipline education coupled with an internship/apprenticeship-based training.

Keywords: Library, Archive, Museum (LAM) Convergence, Data curation, Digital library, Library and Information Science (LIS) Education

Introduction

The discussion of the union of education of information professional in archives, museums, and libraries is beginning. This past year the Governing Board of the International Federation of Library Associations and Institutions (IFLA) established a working group to explore the convergence of education of information professionals for positions in archives, museums, and libraries. While evidence of such convergence was found in North America and in some other places internationally, clearly there is considerable additional research and discussion that is needed to determine the most effective way to proceed to realize a convergence in what is now a very divergent area of identifying educational goals and competencies in these three broad institutions that employ information professionals.

Goals or Outcomes

The focus of education of information professionals in iSchools on digital collections makes the exploration of the potential for iSchools to be a focal point for converging information professional education for the disciplines of Library Science, Archival, and Museum Studies especially relevant. The
fact that many iSchools also contain or have close relationships with schools of Computer Science further supports the potential benefits of iSchool involvement in the exploration of education convergence. One of the specializations relevant to all these disciplines is Digital Curation and this specialization may be especially important to the successful convergence of the education programs. Digital Curation along with other specializations in iSchools may provide the necessary link to achieve convergence. The literature on the management of digital collections, and especially the literature on Digital Curation offers a variety of definitions of the area of study due to the different approaches in each of these disciplines. There are also variations in the curriculum for Digital Curation in North America and in Europe, as well as other parts of the world. These differences need to be identified and harmonized.

The main goal of the workshop is to achieve understanding of the main trends in the interdisciplinary and international research in managing digital collections and related areas within iSchools and how iSchools may design and define more effective curricula for education specifically in the fields of Archive, Museum, and Library and Information Studies.

One possible outcome of the workshop would be the development of a wiki and/or a discussion forum on competencies needed by information professionals in these three fields related to digital collections. Such a communication channel might possibly be supported/sponsored by international organizations such as IFLA (The International Federation of Library Associations and Institutions).

Relevance to the Conference

Most of the participants in the iSchool Conference are involved both in Education and Research, and many of them are to some extent involved or interested in Digital Curation. The workshop intends to provide an opportunity for discussing and sharing experiences both in research and education contexts.

Proposed Program

9:00-10:00

Introduction to topic: Three invited speakers provide twenty minute introductions to the current state of education in archives, museum, and library education with a focus on the discussion of where convergence of the educational programs might be successfully obtained.

10:00 – 10:30

Three invited speakers from iSchools participate in a panel to discuss how their current education programs might contribute to the possible convergence of the information competencies needed in these three information institutions.

10:30 – 11:00

Break

11:00 – 12:00

The workshop participants will be broken into groups (four or five groups of about six participants) based on specific interests in research and development of teaching courses in one of the three areas targeted for convergence in education and will assess the challenges and best strategies for obtaining a successful convergence of information professional education as it relates to working with digital collections. The discussion will be facilitated by the Workshop organizers.

12:00 – 12:30

The workshop organizers will report on the results of the small group discussion and will summarize the workshop findings, review possible next steps, and propose specific ways the participants can continue to communicate and move forward on the desired goals identified during the workshop.
Special Workshop on Information Privacy

Alessandro Acquisti
Carnegie Mellon University
acquisti@andrew.cmu.edu

Deirdre Kathleen Mulligan
UC Berkeley
dkm@ischool.berkeley.edu

Abstract

The Special Workshop on Information Privacy will consist of a one-day presentation of high-quality research in the area of information privacy. It will include 1) invited papers by scholars from participating schools and 2) submitted papers. The day will include four sessions, each of which will feature two paper presentations by authors and ample time for discussion facilitated by a designated commentator.

Our goal is to enhance ties between scholars within the iSchool communities researching privacy and related topics. iSchool faculty, alumni, and students consistently make important contributions to research and policy developments around privacy. This special workshop seeks to facilitate dialogue between different parts of the privacy community, support its continued growth, and identify areas for potential collaboration.

The following papers were selected for the workshop:

Peer-produced Privacy Protection: A Common-pools Approach
Vaibhav Garg, Sameer Patil, Apu Kapadia, L Jean Camp

Abstract: Privacy risks have been addressed through technical solutions (e.g., privacy-enhancing technologies) as well as regulatory measures (e.g., Do Not Track). These approaches are inherently limited as they are grounded in the paradigm of a rational end user who can determine, articulate, and manage his or her consistent privacy preferences. An implication of the rational end user paradigm is that self serving efforts to implement individual privacy preferences lead to socially optimal outcomes regarding information sharing. As a result, solutions to specific privacy risks are developed, and even mandated, without effective reduction in overall harm. We present a systematic framework to examine the limitations of current technical and policy solutions. To address the shortcomings, we argue for considering information sharing to be transactions within a community. Outcomes of privacy management can be improved at a lower overall cost if peers -- as a community -- are empowered by appropriate technical and policy mechanisms. Design for a community requires encouraging dialogue, enabling transparency, and supporting enforcement of community norms. We provide examples of technical design and policy initiatives that leverage such commons based communal governance through peer production of privacy protection.

Theorizing Privacy’s Contestability: A Multi-Dimensional Analytic of Privacy
Deirdre K. Mulligan, Colin Koopman

Abstract: The concept of privacy, despite its centrality for contemporary liberal democratic cultures, remains remarkably contested. The scholarly literatures on privacy theory, privacy law, and the morality of privacy present a dizzying array of diverging conceptualizations and analyses of privacy. This contestability of privacy has been widely observed by privacy scholars. A common thread running throughout these observations is that privacy is polysemous not only in the context of theoretical disputes about privacy’s meaning, but also in the context of privacy’s everyday use.
This article argues that privacy's ambiguity reflects its keen social importance—not as some have suggested a lack of due attention or emptiness. Privacy is an "essentially-contested concept": its value and power—its ability to do work in the world—derives, in part, from its ambiguity. Like other essentially-contested concepts, privacy cuts so much to the heart of our moral, political, legal, and cultural self-understandings that we cannot but disagree over its meaning, application, implementation, and justification. The attendant debates surrounding privacy, while often frustrating, evidence productive negotiations over meaning in rapidly-changing social contexts. Yet, the disagreement at both the theoretical and practical levels over privacy’s meaning and purpose has intensely troubling practical consequences. It can stymie action as an assumed commonality of purpose gives way to an underlying plethora of meanings with separate goals, privacy is decreed too fickle and indeterminate to be advanced through collective decisions in legislative, regulatory, and scientific fora. It can leave wrongs experienced as privacy violations orphaned as theories of privacy fail to connect with human experience. Bandied about but never richly mapped, privacy becomes easy to devalue and less likely to organically evolve and extend. Ambiguity becomes an excuse for disregarding privacy claims—despite visceral and broad appeal, and vociferous support. The challenge is to bring the strength of privacy—its rhetorical power and supple and polysemous nature—into action.

The core contribution of this article is a multi-dimensional analytic of privacy that facilitates a richer analysis of operative concepts of privacy. This analytic mapping provides a toolkit for performing anatomies of privacy as experienced. The analytic delineates a range of dimensions across which concepts of privacy vary including: objects, justifications, exemplars, targets, subjects, actions, offenders, mechanisms, providers, contexts of practice, and scope. In the article’s penultimate section, we use our privacy analytic to illuminate a set of three prominent privacy cases. These examples reveal the value of a more rigorous approach to privacy analysis in legislative drafting, technical design, and argumentation. In the final section, we return to our central argument that privacy’s essential contestability is key to its ongoing relevance and utility in political and social life but only under the guidance of detailed analytical tools that enable us to discern the contours of our many privacies, map them onto contexts, and construct mechanisms to protect them.

The Impact of Privacy Regulation on Technology Adoption: The Case of Health Information Exchanges

Idris Adjerid, Alessandro Acquisti, Rema Padman, Rahul Telang, Julia Adler-Milstein

Abstract: Health Information Exchanges (HIEs) are innovative healthcare technology initiatives that increase coordination between healthcare providers. Their purpose is to improve efficiency and quality of care through enhanced sharing of patient data. To soothe privacy concerns associated with HIE development, however, numerous states have enacted laws establishing strict patient consent requirements for medical data shared through HIEs. We investigate the impact of privacy consent regulation on the adoption and success of HIEs. We find that among all states with laws intended to promote HIE adoption, those that had requirements for patient consent experienced greater HIE adoption and success, while also reporting lower levels of privacy concerns. These findings contribute to the debate over the impact of privacy regulation on technological progress, and provide insights on the delicate balance between privacy concerns and the benefits of technology adoption.


Alan Rubel

Abstract: Public and research libraries have for some time been providing an increasing proportion of resources in electronic formats. The tension between the provision of electronic resources and patron privacy is widely recognized in the LIS literature. But how to assess trade-offs between patron privacy and access to electronic resources remains elusive. One reason is that there is a conceptual problem regarding the nature of intellectual freedom. The gist of the issue is this: Traditionally, the LIS literature and the library profession have understood patron privacy as a facet of intellectual freedom, and it is plausible that there is some relation between information privacy and liberty or freedom (I use these
interchangeably). And while certain types of electronic resource use may diminish patron privacy, thereby diminishing intellectual freedom, the opportunities created by such resources also appear liberty-enhancing. Any attempt to adjudicate between privacy loss and enhanced opportunities on intellectual freedom grounds must therefore provide an account of intellectual freedom capable of addressing both privacy and opportunity. That’s my task here. I will argue that intellectual freedom is best understood as a form of positive freedom, where a person’s freedom is a function of the quality of her agency. Using quality of a person’s agency as the lodestar, I set forth several principles for assessing adoption of electronic resources and privacy protections.

Patron Privacy in the “2.0” Era: Avoiding the Faustian Bargain of Library 2.0
Michael Zimmer

Abstract: As libraries begin to embrace Web 2.0 technologies to serve patrons—ushering in the era of Library 2.0—unique dilemmas arise in the realm of information ethics, especially regarding patron privacy. The norms of Web 2.0 promote the open sharing of information—often personal information—and the design of many Library 2.0 services capitalize on access to patron information and might require additional tracking, collection, and aggregation of patron activities. Thus, embracing Library 2.0 potentially threatens the traditional ethics of librarianship, where protecting patron privacy and intellectual freedom has been held paramount. The question is not whether libraries will move towards Library 2.0, but how they will do it, and whether they can preserve the contextual integrity of patron privacy and maintain their professional librarian ethic, while also providing enhanced services to their patrons. This article will provide an ethical examination of the emergence of new Library 2.0 tools and technologies in relation to existing ethical norms of information flow within the library context. By doing so, librarians and information professionals will be better situated to avoid—or at least renegotiate—the impending Faustian bargain regarding patron privacy in the “2.0” era.

A Retreat from the Panoptic: One Public Library’s Experience with Video Surveillance
Bryce Clayton Newell and David P. Randall

Abstract: This paper presents the findings of a qualitative case study examining why one public library installed video surveillance systems and then later reversed course and completely removed the previously installed systems. We found that the library initially installed the system as a response to specific incidents of crime without central administrative oversight, and that the removal was prompted by deteriorating relationships with local police departments over the library’s position that the video footage was exempt from public disclosure under the state’s library records privacy law. The library system subsequently removed all of their cameras in 2011, claiming the cameras were not in sync with library commitments to intellectual freedom and patron privacy, despite the fact that library staff expressed strong interest in retaining the cameras and were concerned about staff safety and crime prevention. We also found evidence of surveillance creep.

OTP-PAKE: A More Secure Alternative to HTTPS
Robert Huijie Deng, Divyan Munirathnam Konidala, Yingjiu Li, Yongdong Wu

Abstract: The HTTPS protocol is widely used for secure communications. However, the HTTPS approach subjects to various attacks which have serious implications on security and privacy. This paper shines light on the seldom used and yet cryptographically strong password-authenticated key exchange (PAKE) protocols to secure the communication channel between a client and a server. We justify that PAKE-based protocols offer better security and privacy protection than the HTTPS paradigm. We point out a particular shortcoming in the PAKE-based protocol and propose the use of one-time password (OTP) along with PAKE in order to overcome this shortcoming. We also offer interesting discussions with regard to OTP-PAKE protocol and propose a simple Diffie-Hellman key exchange based OTP-PAKE protocol design.
Silent Listeners: The Evolution of Privacy and Disclosure on Facebook
Fred Stutzman, Ralph Gross, Alessandro Acquisti

Abstract: Over the past decade, social network sites have experienced dramatic growth in popularity. Once a youth-centric phenomena, the use of services like Facebook and Google+ now cross almost all demographics. With this growth, users have been challenged to manage novel privacy hurdles and negotiate nuanced balances between withholding and disclosing personal information. To date, however, no study has documented the evolution of disclosure and privacy behavior on popular social networking sites over an extended period of time. In this manuscript we use actual profile data from a longitudinal panel of 5,076 Facebook users to understand how their disclosure behavior changed between 2005 and 2011. Our findings suggest that 1) Facebook users in our dataset became increasingly protective of their personal information over time, limiting data shared with strangers; 2) changes effected by Facebook over the period of time under our observation, such as interface changes, arrested or in some cases inverted that trend; 3) over time, the amount and scope of personal information that Facebook users reveal to connected “friends” profiles has increased dramatically—but so have the disclosures to “silent listeners” on the network: third-party apps, Facebook itself, and (indirectly) advertisers.

Profiling the Profilers: Deep Packet Inspection for Behavioral Advertising in Europe and the United States
Andreas Kuehn and Milton Mueller

Abstract: This paper examines the use of deep packet inspection (DPI) in online advertising, and analyzes the effects public pressure, regulatory actions and judicial and policy-making proceedings had on those deployments. DPI, which allows Internet service providers (ISPs) to monitor the content of data packets in real-time, can be considered a disruptive technology because of the way its use conflicts with pre-established principles and norms of Internet governance.

In this comparative study, we examine the rise and fall of NebuAd in the U.S. and Phorm in the United Kingdom and the European Union. We also include some less visible companies and spill-overs to Brazil and South Korea. We conduct a comprehensive analysis of these cases – from the early development and secret trials of the technology to the regulatory actions, business failures and litigation in the aftermath. Looking at a timeline of several years that covers the dynamic technical, economic and institutional interactions at play, the framework contrasts distinct actors, actor constellations and modes of interaction across institutional settings to illustrate similar and divergent policy outcomes. This research is based upon comprehensive analysis of political and legal documents and a series of interviews with DPI vendors, Internet advocates, engineers, and advertisers.

The narrative follows four stages that we have found repeatedly in similar case studies of DPI deployments: 1) secret deployment, 2) uncontrolled public disclosure of the deployment, 3) civil activism around net neutrality and privacy norms, 4) political, legal and regulatory proceedings to resolve the conflicts. This framework highlights the interaction of technical, economic and institutional factors that are at work when politically contested technologies with a disruptive potential are deployed on the Internet. In this case, as in many others, the analysis shows how the deployments ran afoul of established principles and expectations, how the “notification” and “consent” practices so crucial to privacy law failed to bridge the gap between the expectations of Internet users and the formal legal definition applied by the courts, and how this gap led to intense political pressure and market exit of DPI-based advertising platforms in both countries.
Educating Information Professionals for 2050

Gary Marchionini  
University of North Carolina  
gary@ils.unc.edu

Mike Eisenberg  
University of Washington  
mb@uw.edu

Liz Liddy  
Syracuse University  
liddy@syr.edu

Cathy Marshall  
Microsoft  
cathymar@microsoft.com

Abstract

Students who graduate from I-Schools in 2013 will very likely be working in 2050. We cannot predict what the world will be like in 2050 but we should imagine what kind of world we would like to see in 2050. To this end, in June 2012 we held a 2-day symposium and conference in Chapel Hill called Information Professionals 2050 that brought together 16 thought leaders to discuss what kind of information future we would like to see and consider how we best prepare information students today for careers that will still be active at mid-21st century. The proceedings from the symposium and conference are available at http://sils.unc.edu/sites/default/files/publications/Information-Professionals-2050.pdf. This panel brings together three of those thought leaders and the organizer of the meetings. Mike Eisenberg will discuss education trends such as demassification and the evolution of personal education records and how technologies are bringing fundamental changes in how, when, and where learning takes place. Liz Liddy will discuss how entrepreneurship is permeating I-Schools and adding rich experiences to learning, agile approaches to instruction, and new partnerships with industry. Cathy Marshall will discuss the potentials and perils of big data based on ethnographic investigations of how personal and corporate data are collected and used. Gary Marchionini will discuss the core values that information professionals share and how these values drive I-School approaches to socio-technical challenges and education.

Keywords: education, information trends


Copyright is held by the authors.
Abstract

This alternative event for the 2013 iConference is a combination of lightning talks, a demonstration of an assessment technology for knowledge construction in complex domains, and a hands-on exercise in using the tools discussed. The unifying logic for this presentation is that meaningful learning often involves solving challenging and complex problems that allow for multiple solution approaches and a variety of acceptable solutions. While it is important to prepare students to solve such problems, it is difficult to determine the extent to which various interventions and programs are contributing to the development of appropriate problem-solving strategies and attitudes. Simply testing domain knowledge or the ability to solve simple, single-solution problems may not provide support for improving individual student ability or relevant programs and activities. A reliable and robust methodology for assessing the relevant knowledge constructions of students engaged in solving challenging problems is needed, and that is our focus.

Keywords: complex problems, dynamic assessment, information representation, measurement

Introduction

The objectives of this special event session are (1) to build awareness of the connections between the representation of information and the construction of knowledge, (2) to illustrate a framework for conceptualizing these relationships, (3) to introduce the issue of measuring how information representations contribute to knowledge construction, (4) to demonstrate a specific technology for assessing how knowledge is being constructed by a learner, and (5) to provide an opportunity to practice with regard to a common information representation/knowledge construction problem.

The practical knowledge construction problem to be illustrated and involved in the practical activity is a design problem. Designing and developing learning resources and activities is an experience familiar to most educators. The specific design problem involves a decision about multiple ways to represent information about a particular complex and ill-structured problem-solving situation to be presented to learners so as to promote deep thinking about that complex problem. The topic for the design exercise is climate change. The targeted learners are 9th grade students in an earth science class. The problem to be introduced involves developing a policy with regard to carbon dioxide (CO2) emissions aimed at reducing global warming, which is presented mid-way through the course and used as a capstone activity to be completed by small groups (3 to 5) of students. A representative expert solution will be available for discussion towards the end of the session to discuss the notion of dynamic formative feedback based on the knowledge constructions provided by participants.

Discussion

The session includes short, provocative, lightning talks (Pechu Kucha) by each of the panelists. The focus of these lightning talks is on methods, tools, and technologies that can be used to assess how the use of information in a learning situation is contributing to the construction of knowledge. The first
issue is a general framework that involves the introduction of representations of mental models and how they can be used as a basis for assessment; a tool that has been used successfully for this purpose is HIMATT (Highly Integrated Model-based Assessment Tools and Technologies; Pirnay-Dummer, Ifenthaler, & Spector, 2010). An extension of HIMATT that explicitly supports dynamic formative feedback to learners engaged in problem solving activities is the second issue addressed (Ifenthaler, 2012). A simulation-based learning environment for pre-service teachers and how it can support refinement of classroom teaching is then presented (Knezek & Christensen, 2009). The focus then shifts to special populations and how these tools and technologies can be used to support individual needs (Tyler-Wood, Ellison, Lim, & Periathinuvadi, 2011). Motivational and affective factors (e.g., enjoyment) that can be considered in the implementation of the tools and technologies is also discussed (Kim, 2012). The concluding provocative contribution demonstrates a synthesis of the panelists’ talks to understand individual mental model representations and support the effective construction of knowledge.

The basic logic behind these presentations begins with a widely accepted constructivist epistemology. Individuals create internal representations to make sense of their experiences. These internal representations are not directly observable but they are critical for effective knowledge construction and problem solving. It is possible to elicit re-representations of these constructions during a problem solving activity and use those to (a) identify student problems, including both cognitive and non-cognitive problems, (b) offer meaningful, just-in-need feedback, and (c) evaluate the efficacy of a learning activity. The underlying assumption is that measuring progress of learning and understanding is fundamental for promoting effective knowledge construction and deep understanding of complex and challenging problems.

At the end of these presentations, there are three additional activities: (a) reactions and reflections from participants, (b) a structured exercise for pairs of participants using a particular annotated concept mapping technique to construct a knowledge model about a selected topic of general interest to information scientists and technologists, and (c) opportunities to share some of these constructions with the full group.

**Conclusion**

The issue of measuring the impact of information-centered tools, technologies and methodologies on the effective construction of knowledge is of critical importance to the general education community and of specific relevance to those working in the area of information studies. The way that information is represented and used in educational contexts is known to have an effect on how different persons [mis-]interpret and [mis]-understand that information. Moreover, the ability to offer real-time, problem- and learner-specific feedback on representations and understanding is of obvious relevance in helping learners construct effective mental models and knowledge structures. The goal of this session is to bring these issues to the forefront and to illustrate at least one approach that has shown strong promise in promoting effective knowledge construction based on information representations. The expected short-term outcome of this session is a paper indicating tools and technologies that appear useful in promoting effective knowledge construction and problem-solving skills.

**References**


The Problem of Data: iSchools and the New Data Professional

Spencer D. C. Keralis  
University of North Texas  
Digital Scholarship Co-Operative  
spencer.keralis@unt.edu

Shannon Stark  
University of North Texas  
Libraries  
shannon.stark@unt.edu

Abstract

In this fishbowl event, we invite iSchool and LIS faculty and administrators, recent graduates, Library staff and administrators, researchers, and data professionals to speak to the state of data curation education. What are iSchools doing to train data professionals? What are institutions doing to address the professional development needs of librarians? What are the real skills libraries should be demanding of LIS graduates? And how can the various constituents in the data curation lifecycle collaborate to develop and foster these skills in emerging professionals?

Keywords: Data Curation, Data Lifecycle, Pedagogy, Professional Development, Library Education

Introduction

In the Council on Library and Information Resources (CLIR) report The Problem of Data (Asher, Keralis, & Jahnke. Washington, D.C.: CLIR, 2012), researchers concluded that while iSchools and the LIS field perceive data curation as part of their domain expertise, in the current curricular landscape of these programs, it is improbable that most students would “encounter a data-centric course in their line of study” (Keralis, 2012). Further, the recent Data Conservancy’s Data Curation Curriculum Search underscored the paucity of truly data-centric courses, and the difficulty in identifying courses that might provide data management skills in most U.S. LIS programs. Given these circumstances, LIS students must be “committed to crafting a data-intensive education for themselves … in order to come out of most existing U.S. LIS programs with the skills and knowledge necessary to support the needs of data-intensive research” (Keralis, 2012, 34-35). Further, it is clear from the CLIR study that there is a gap between how academic libraries understand their place in the research process and how libraries are perceived by researchers: “Few researchers see the library as a partner … and most of the researchers regard the library as a dispensary of goods rather than a locus for badly needed support” (Asher & Jahnke, 2012, 16).

Audience

The purpose of this event is to generate conversation around the question of how best to develop skills for data professionals. The audience for this event is anyone with a stake in the data curation lifecycle: representatives from funding agencies; LIS and iSchool students, faculty, and administration; librarians; library administrators; training vendors; technologists; researchers; graduate students in the sciences and social sciences; and science and social science faculty.

Acknowledgements: Research contributing to this alternative event was funded by the Alfred P. Sloan Foundation in an award to the Council on Library & Information Resources/Digital Library Federation, and by a Laura Bush 21st Century Librarians Grant from the Institute of Museum and Library Services, awarded to the University of North Texas Libraries and College of Information for the DataRes Project (http://datamanagement.unt.edu). We are grateful to Martin Halbert, Dean, UNT Libraries; William E. Moen, Associate Dean for Research, UNT College of Information; Elliott Shore, Chief Information Officer, Bryn Mawr College; and Rachel Frick, Director, DLF for their commitment to advancing education efforts for the next generation of data professionals in academic libraries.


Copyright is held by the authors.
Activities

Supporting documentation and resources designed to stimulate conversation will be available at http://datamanagement.unt.edu/fishbowl.

The event will consist of a brief introduction to the issues surrounding educational support for data professionals; the remainder of the event will be conducted as a fishbowl, in which audience members will come forward and speak about their perspective on the issues, in response to key questions framed by the presenters. The presenters will record these responses, and help facilitate dialogue among the attendees. Key findings from the conversation will be shared after the event at event website.

Relevance

The event will appeal to conference attendees in terms of content, because data curation is having an ongoing and profound impact on iSchools, those who receive their training in iSchools, and those who employ iSchool graduates. The format allows a more democratic range of expression for attendees to relate information and challenge premise of the event, and allows more people to be heard than a traditional panel or plenary event.

Logistics

The event will be 90 minutes: 10 minute introduction, 70 minutes of conversation, and 10 minutes of wrap up. The event would be optimal with approximately 50-75 attendees. The room should be arranged in a half-circle with a chair or podium and microphone at the center.

References

Data Curation Curriculum Search. Available at http://cirssweb.lis.illinois.edu/DCCourseScan1/index.html
Calling All Computer Scientists and Social Scientists: Establishing a Research Agenda for Computational Social Science

Kenneth R. Fleischmann  
University of Texas at Austin  
kfleisch@ischool.utexas.edu

Sean P. Goggins  
Drexel University  
spg46@drexel.edu

James Howison  
University of Texas at Austin  
howison@ischool.utexas.edu

Matthew Lease  
University of Texas at Austin  
ml@ischool.utexas.edu

Douglas W. Oard  
University of Maryland  
oard@umd.edu

Abstract

Computer scientists and social scientists are major constituent groups in most iSchools. Our goal is to bring these two groups together, including social scientists who study social phenomena broadly or deeply and computer scientists who have computational approaches that can be used to balance the trade-off between depth and breadth (Cheng, Fleischmann, Wang, & Oard, 2008; Fleischmann et al., 2012; Fleischmann, Oard, Cheng, Wang, & Ishita, 2009; Fleischmann, Templeton, & Boyd-Graber, 2011; Oard, 2009). Some researchers have already begun to define themselves as computational social scientists (e.g., Cioffi-Revilla, 2010; Gilbert, 2009; Lazer et al., 2010). This event will be highly relevant to this group, however, and it will also be useful to computer scientists and social scientists who are just beginning to consider collaboration across this disciplinary divide.

Broadly, computational social science can be seen as the application of computational approaches (including the development of new computational approaches) to systematically study social phenomena. The first wave of computational social science focused on agent-based modeling (e.g., Cioffi-Revilla, 2010; Gilbert, 2009). The second wave of computational social science involved social network analysis (e.g., Lazer et al., 2010; Mascaro, Novak, & Goggins, 2012). The third wave of computational social science is automatic content analysis, which employs natural language processing techniques to scale up content analysis (e.g., Ishita, Oard, Fleischmann, Cheng, & Templeton, 2010; Templeton, Fleischmann, & Boyd-Graber, 2011a, 2011b; Zhou, Fleischmann, & Wallace, 2010). iSchools contain researchers from all three waves of computational social science. iSchools can also play a leading role in developing future waves of computational social science.

One goal of this event is to bring together researchers with complementary interests in computer science and/or social science. Another goal is to help define the emerging field of computational social science and propose some specific research that would help researchers to capitalize on this opportunity. One more goal is to talk about obstacles to successful computational social science research and opportunities to overcome these obstacles, such as new funding mechanisms for interdisciplinary research and collaboration guidelines for collaborating across such broad disciplinary divides. Yet another goal is to set the stage for a larger workshop in the future, perhaps at a future iConference. The final goal is to ensure that the iSchools continue to play a leadership role in the development of computational social science – iSchoolers are already at the cutting edge of this emerging field, and iSchools are uniquely situated to lead this field as interdisciplinary homes to both computer scientists and social scientists.

The event will begin with a brief introductory presentation. We will then hold four rounds of small group lightning discussions, with the requirement that each discussion group must contain at least one computer scientist and at least one social scientist. We will have a list of questions for each round of lighting discussion to cover. We will end with a full-room discussion with the remaining time.
The topic that we will explore, computational social science, should appeal to a wide range of attendees, spanning computer scientists and social scientists. We hope that this event will help people to realize that they are building and to commit to building the field and to make the iSchools central within the field. We also hope that the dynamic format and the continuity of at-conference and online interactions will provide a change of pace from more traditional conference events typically found at other conferences.

*Keywords:* computational social science, agent-based modeling, social network analysis, automatic content analysis, natural language processing

References


Experiencing Science in Informal Learning Environments: Tales from the Field

Mega Subramaniam
University of Maryland
mmsubram@umd.edu

June Ahn
University of Maryland
juneahn@umd.edu

Nicole D. Alemanne
Florida State University
nda06c@my.fsu.edu

Mike Eisenberg
University of Washington
mbe@uw.edu

Sean Fullerton
University of Washington
sfullert@uw.edu

Michelle M. Kazmer
Florida State University
mkazmer@fsu.edu

Marcia Mardis
Florida State University
marcia.mardis@cci.fsu.edu

Abstract

Informal learning environments such as libraries and museums can play a vital role in attracting youth and children to science learning. Libraries and museums offer potential opportunities for children and youth to explore science in creative ways that are freed from the formal identities and stigmas attached to the academic classroom, by engaging in technology, media, information and pedagogy that is appealing to them. Research on the potential of these information institutions is currently still in its infancy, and a handful of researchers in the information field are exploring such opportunities. Information institutions can offer several unique benefits to science learning -- engaging young people in inquiry, which includes vital information literacy skills, providing access to diverse resources and materials, connecting to young people's personal interests, and making science relevant to their lives. In this event, we will share stories from innovative research that highlight how information institutions offer the above-mentioned benefits to science learning.

Keywords: informal learning, science learning, digital media, socio-cultural, participatory culture

Purpose and Intended Audience

We present a participatory event that will involve researchers who will share stories and artifacts of successful learning environments in libraries and museums that encourage scientific inquiry, appreciate diversity, connect to the interests of youth and children and encourage their socialization. The stories and artifacts (which may include zines, digital journals, websites) are drawn from ongoing research that the authors are leading. Audiences will be encouraged to share their own stories as they relate to the thematic presentation of the unique roles of information institutions. We anticipate that this session will be appealing to researchers who are involved in scholarly work related to digital media, participatory culture, literacy development, design of technologies, scientific inquiry, K-12 education, youth and children services in libraries, museums informatics, and informal learning among youth and children.

Agenda

Mega Subramaniam will lead the session by introducing the unique roles that information institutions play in science learning via stories and/or video presentations. Researchers will share lively narration of experiences and artifacts produced by youth and children from their respective research
projects, as they relate to each unique role of information institutions being discussed. As they share stories and artifacts, the researchers will highlight the role that information professionals play in creating and sustaining successful science learning environments, the role that technology and media plays in such environments, the mastery of science and participatory practices, and the intersection of all these components in the design of socio-scientific learning environments for youth and children.

Mega Subramaniam and June Ahn will share stories and artifacts from their research on Sci-dentity <http://scidentity.umd.edu>, a project in which researchers are collaborating with school librarians to design ways to incorporate science storytelling, new-media literacies, and participatory culture to ignite students’ interest in science. Sci-dentity is an after-school program that encourages reading of science fiction, popular fiction, and graphic novels, watching sci-fi movies, and playing science-infused games among middle school youth. In this program, youth imagine the underlying science that inspires these popular forms of media, create their own science-inspired stories, and write and share these stories on a private social media site. This research is funded by the National Science Foundation.

Nicole D. Alemanne and Michelle M. Kazmer will share stories and artifacts from the Habitat Tracker project <http://tracker.cci.fsu.edu>, which is providing teachers with better support for linking field trips to a wildlife center with a standards-based, scientific inquiry curriculum while engaging fourth and fifth grade children with their own science education. Using an iPad application and a website, children interact with digital journals, collaboration and analysis tools, and a shared online observations database to develop, refine, and answer their own scientific research questions. Almost 1,400 children in 25 schools have participated in Habitat Tracker during the three-year project, recording over 4,600 habitat, animal, and weather observations to-date. This research is funded by the Institute of Education Sciences.

Mike Eisenberg and Sean Fullerton will share about the early development of the Personal Education Record (PER) system, which they are investigating as a way to allow young people to document learning experiences in museums, libraries, virtual spaces, and other places where learning occurs outside of school. The PER is designed to provide a well-organized system to connect young peoples’ interests and learning experiences outside of school to a framework that allows young people to record these experiences and show how the various activities fit within broader curriculum areas such as biology or physical science. Our hope is that the PER will allow young people to see how interest-driven, but sometimes isolated learning experiences fit within a more cohesive curricular framework like those traditionally provided by courses and textbooks.

We will conclude with an interactive conversation with the audience (led by Marcia Mardis) on the potential of innovative research in creating science learning environments in libraries and museums. The discussion will be centered on the types of research problems that can be pursued, and the unique contribution that library and information science scholars can offer to K-12 science learning.

References

Whither the Child? The iSchool Approach to Research Relating to Children and Adolescents

Leanne Bowler  
University of Pittsburgh  
lbowler@sis.pitt.edu

Kate McDowell  
University of Illinois  
kmcdowel@illinois.edu

Denise Agosto  
Drexel University  
denise.agosto@ischool.drexel.edu

Abstract

This alternative event will investigate how iSchools are approaching research in the area of children and adolescents. Participants will work together to tackle this overarching question: Where does youth research fit in to the iSchool movement? Participants will be asked to contribute to the development of a collaborative agenda for youth-related research in the iSchools by identifying related challenges and opportunities within the iSchool community.

Keywords: iSchools, children, adolescents, research agenda, research methods

Introduction: This alternative event will investigate how iSchools are approaching research in the area of children and adolescents. Participants will work together to tackle this overarching question: Where does youth research fit in to the iSchool movement?

The iSchool movement grew, in part, out of the traditional library school. This is especially true of schools that have a history of preparing students to work with children and adolescents in public and school libraries. In fact, many of the iSchools are ALA accredited and continue to provide courses to train children’s, young adult, and school librarians. The web site for the iSchool caucus identifies eight schools with a specialization in children and adolescents, including school librarianship (Drexel, Illinois at Urbana-Champaign, Indiana, Michigan, Pittsburgh, Syracuse, Texas and Washington). A survey of the web sites for each of the 38 iSchools in the iCaucus found at least 42 faculty members from 22 iSchools who teach or conduct research with a focus on children and adolescents. When we looked at the proceedings for the 2011 and 2012 iConferences, we saw evidence of contributions to research, with five of 177 papers in 2011 and nine of 148 papers in 2012 on topics related to children and adolescents.

The iSchools have a broad vision of information, people, and technology, and they focus on critical information issues in society. The iSchool movement has allowed the lens of inquiry to widen, opening up new areas for research. But will this new vision continue to include research related to children and adolescents? How does the language around youth research (qualitative, outcome-based, story, etc.) mesh with the language of “information”? Are there opportunities waiting to be discovered? Or have we left the children behind?

Purpose: The purpose of this alternative event is to:

- Open up a conversation about the nature of research in the area of children and adolescents at iSchools;
- Identify opportunities and barriers for researchers conducting research in the area of children and adolescents at an iSchool;
- Bring together researchers, educators, students, and others interested in issues relating to youth and information;
- Begin to build lasting youth-focused communities and support systems within the iSchool movement.

Copyright is held by the authors.
Intended Audience:

- Professors who conduct research in the area of children and youth;
- Doctoral students;
- Prospective doctoral students;
- Deans and directors of iSchools;
- Librarians;
- Designers of children’s media.

Length of time: 90 minutes

Agenda: Participants will be asked to contribute to the development of a collaborative agenda for youth-related research in the iSchools by identifying related challenges and opportunities within the iSchool community. Questions for participants will include:

1) What is the role of youth-related research in your iSchools today?
2) What should the role of this research be in the future?
3) What are some of the unique challenges, barriers, and opportunities for youth-related research?
4) Where should our community of youth-related researchers go from here?

Relevance to the Conference/Significance to the Field: This event will build on traditional strengths of the field while inviting new approaches and vision to research and teaching related to children and adolescents. It will focus on youth-related scholarship in action, with an emphasis on identifying the kinds of youth-related data gathering, innovations in youth-related research, and knowledge of youth and information that is unique to the iSchool community.
Values Interventions: Ethics Scholarship in Action

Katie Shilton  
University of Maryland-College Park  
kshilton@umd.edu

Michael Zimmer  
University of Wisconsin-Milwaukee  
zimmerm@uwm.edu

Finn Brunton  
University of Michigan  
finnb@umich.edu

Deirdre Mulligan  
University of California-Berkeley  
dkm@ischool.berkeley.edu

Abstract

The Internet has permeated the economic, political, cultural and social domains of global society and transformed the way in which we present and transmit knowledge. The infrastructure underlying the Internet continues to evolve, with ramifications for not only the technical protocols that govern the way the Internet functions, but also implications for social, economic, and legal issues. Internet protocols affect debates about intellectual property, privacy, security, and information access. What values (e.g. privacy, trust, security, equity, transparency, etc.) are embedded in our current Internet architecture, and what will be considered in the design of future Internet architectures? This panel describes an ongoing project to put ethics scholarship in action: a values intervention into the design of future Internet architectures.

Panelists from the NSF-funded Future Internet Architecture Values in Design Council will speak about the effort to introduce social and ethical discussion into technical design practice, and the methods and challenges of such values interventions.

Keywords: ethics, values in design, Internet architecture

Panel Description

The Internet has permeated the economic, political, cultural and social domains of global society and transformed the way in which we present and transmit knowledge. The infrastructure underlying the Internet continues to evolve, with ramifications for not only the technical protocols that govern the way the Internet functions, but also implications for social, economic, and legal issues. Internet protocols affect debates about intellectual property, privacy, security, and information access. What values (e.g. privacy, trust, security, equity, transparency, etc.) are embedded in our current Internet architecture, and what will be considered in the design of future Internet architectures? This panel describes an ongoing project to put ethics scholarship in action: a values intervention into the design of future Internet architectures.

Panelists from the NSF’s Future Internet Architecture Values in Design Council (http://www.nyu.edu/projects/nissenbaum/vid_council.html) will speak about the effort to introduce social and ethical discussion into technical design practice, and the methods and challenges of such values interventions.

The intersection of information systems and social values is an important question facing policy scholars, social scientists, and engineers. The design of technology is never value-neutral, and questions of what, and whose, values are embodied in software and system architecture have been controversial for decades (Alsheikh, Rode, & Lindley, 2011; Friedman & Kahn, 1997). Affordances built into a technology may privilege some uses (and users) while marginalizing others, highlighting values as a critical if sometimes invisible influence on the design process. Internet architecture, in particular, carries a number of values questions in its design. In Protocol Politics (2009), DeNardis suggests that there are many ways that networking protocols can change distributions of resources and power. New Internet protocols would impact several political issues, including protection of privacy, access to knowledge, and trust and authenticity of data. They would also have a large impact on existing institutional power struggles. Challenges and social debates like network neutrality (Lemley & Lessig, 2001), wiretapping backdoors (Landau, 2011), and cybersecurity (Clark & Landau, 2011) are all affected by Internet protocols.

Acknowledgement: The Values in Design Council is supported by NSF Eager: Values in Design in the Future Internet Architecture, CNS/NetS 1058333 (Helen Nissenbaum, Principle Investigator)


Copyright is held by the authors.
Building consideration of social values directly into engineering practice is a growing movement in technology design (Johnson, 2000; Knobel & Bowker, 2011; Manders-Huits & Zimmer, 2009; Miller, Friedman, & Jancke, 2007; Shilton, 2010). Social values are significant characteristics or goals ascribed to by individuals or a group. These are often broadly construed under the categories of, as Friedman et al describe, “fairness, justice, human welfare and virtue” (2006, p. 13), and stem from a variety of ethical perspectives including deontology, utilitarianism, and virtue ethics (Ess, 2009). Because systems perform some tasks better than others, choices about what functions to value are built into tool and platform affordances. Decades of research in engineering ethics, science and technology studies, and information science have demonstrated that values like privacy, consent, or openness can be enabled by a system's features or rules, and are the result of technical choices during the design process (Agre, 1997; Bowker & Star, 2000; Friedman et al., 2006; Friedman & Nissenbaum, 1997).

This panel will describe an interdisciplinary team effort to undertake values-focused interventions with network architects working on Internet protocol redesign as part of the NSF’s Future Internet Architecture (FIA) program. We will discuss social values relevant to Internet design, methods and challenges for intervention in technology design, and solicit audience feedback and perspectives regarding similar efforts for engaging in ethics scholarship in action.

Purpose and Intended Audience

The intended audience is scholars and practitioners interested in the areas of values in design, values-sensitive design, or social science and humanistic interventions into science and engineering practice.

Relevance to the Conference

We still don’t fully understand the emergence of values in technological systems: how affordances that support some values (e.g. efficiency and novelty) over others (e.g. privacy and security) are chosen and implemented during the design process. Design of ubiquitous information systems like the Internet is a complicated, situated process incorporating diverse actors, work processes, and social contexts. This ongoing research seeks to understand how these actors, work practices and contexts contribute to the embedding of particular values in information systems. The iConference is an excellent venue for this conversation, because it brings together researchers and practitioners working across the areas of values in design and values-sensitive design.

Proposed Activities

The event will be organized as an interactive panel discussion. The first 30 minutes will be devoted to an overview of the project and the challenges thus far:

- Michael Zimmer will introduce the Future Internet Architecture projects and the role the Values In Design Council has played as analysts and consultants to the FIA projects, helping to identify junctures in the design process in which values-critical technical decisions arise, locating design parameters and variations that differentially call into play relevant values, operationalizing values to enable transition from values conceptions into design features, and helping FIA designers examine the interplay of technical values embodied in design with respective ethical values.

- Katie Shilton will detail an embedded intervention with one of the five FIA projects: the Named Data Networking project. Shilton's project studies values levers, which are work practices and agents within design that can encourage discussion of values such as privacy, trust, and authenticity during the design process.

- Finn Brunton will discuss the VID Council’s use of “adversarial scenarios” as a means of intervention with the FIA design teams. Asking each design team to address numerous scenarios provided them an incentive to specify how particular features (or changes) within the architecture could feed forward through the options available to actors, and the courses they will take. The scenarios helped raise questions and choices that might otherwise not be apparent within the different projects, and helped make visible various value dimensions not immediately discernable.
• Deirdre Mulligan will discuss particular ethical dimensions of the FIA designs, focusing on the aspects of the system architectures relating to security, trustworthiness, and privacy, as well as commenting on the challenges faced by VID members attempting to engage in this form of ethics scholarship in action.

After presenting details from these participants, we will incorporate audience questions, feedback and discussion for the final 60 minutes of the session.

References


Plan|Play|Pressure|Pause. Engaging Creative Information Practices

Theresa Anderson  
Centre for Creative Practice & Cultural Economy  
University of Technology (UTS), Sydney, Australia  
theresa.anderson@uts.edu.au

Leanne Bowler  
School of Information Sciences  
University of Pittsburgh, PA  
lbowler@sis.pitt.edu

Lisa Nathan  
The iSchool@UBC  
University of British Columbia, Canada  
lisa.nathan@ubc.ca

Eileen Trauth  
College of Information Sciences & Technology  
Pennsylvania State University, University Park, PA  
etrauth@ist.psu.edu

Abstract

This alternative event is one of a series of successful iConference sessions developing a discourse that recognizes and appreciates what the creative milieu described by Howkins (2009) and Florida (2001) means for our creative information practices. Collectively, these events frame a new research area to engage and envision investigations into those practices. The 2013 event critically engages participants in creative practices by drawing upon Anderson’s 4P heuristic (plan, play, pressure, pause) to scaffold creative engagements. Each of these “Ps” represents a different way of engaging with information. Working on the premise that having a mix of the four is critical in nurturing an innovative culture, participants are invited to become more mindful of the mix that might best suit their own contexts: be it in their personal practice or the practice of the communities they wish to serve. To wrap-up, participants envision next steps in this emerging research area.

Keywords: creativity, information practices, innovation cultures, creative ecologies, scholarly practice

Overview

Creativity is a vital part of human experience. In many Western societies we demonstrate our esteem for creativity by attending innovative movies, cutting edge musical performances, avant-garde art galleries, experimental theatre, and valorizing innovative entrepreneurs. We marvel at the creativity evident in inventive engagements with the urban environment (e.g., Knit Graffiti, Parkour, Yarn Bombing). Yet, too often we fail to recognize, appreciate and support the creativity needed to engage with our own ever-changing information environment. Bombarded with deterministic rhetoric that idolizes new information tools, it is easy to gloss over the fact that we creatively appropriate these tools into our daily practice. Until recently, we have not paid enough attention to the critical role creativity plays in relation to information practices.

This alternative event will cultivate budding interest in creative information practices. It is one of a series of successful iConference sessions developing a discourse that recognizes and appreciates what the creative milieu described by Howkins (2009) and Florida (2001) means for our creative information practices. The first creativity focused iConference event took place in 2011 as a preconference workshop. The workshop provided space for participants to take initial steps towards identifying methods for supporting and nurturing creativity. The 2012 iConference hosted an alternative event that focused on infusing creativity into information practices related to the research lifecycle (developing research questions, methodological choices, data analysis, communication research results). The research


Copyright is held by the authors.
process served as an organizing construct for the event. The 2013 iConference event engages participants in creative practices by drawing upon Anderson’s 4P heuristic (plan, play, pressure, pause) to scaffold creative engagements (Anderson, 2012).
Collectively, this series of events frame a new research area, bringing together researchers and practitioners to engage and envision investigations into creative information practices.

Purpose

This alternative event is designed to:
- Identify, articulate, and engage in an array of creative information practices;
- Build a community of researchers interested in developing the field of creative information practices;
- Develop research questions to investigate creative information practices in local and global contexts;
- Generate an agenda for future scholarship and partnerships.

Intended Audience

Consistent with the call for alternative events that might kickstart thinking in creative ways, the engaging content and interactive format of this event is targeted to appeal to attendees interested in discussing, sharing and developing creative information practices. We have crafted versatile activities to accommodate any number of attendees. Our session is designed to enable participants to critically engage in creative information practices.

Details of proposed activities

Our event is framed by Anderson’s 4P heuristic for stimulating creative information practices: plan, play, pressure, and pause (Anderson, 2012). Derived as part of her ongoing ethnographic investigations of research practice and the conditions underpinning creativity and innovation (see for instance: Anderson, 2011), each of these “Ps” represents a different way of engaging with information. Anderson asserts that having a mix of the four is critical in nurturing an innovative culture, although the balance depends on the individual context. During the session participants are invited to become more mindful of the mix that might best suit their own contexts: be it in their personal practice or the practice of the communities they wish to serve.

The event begins with each of the four organizers briefly elaborating on how attributes of the 4P framework are enacted through her personal creative information practice. For example, Eileen Trauth will present and discuss a short video clip showing a scene from her play, iDream, which was written to communicate the results of her research about barriers to women and under-represented minorities in the information technology field. Leanne Bowler will discuss how she enacted the planning piece of creativity through a “creativity audit” that is based on Anderson’s 4P heuristic.

Participants will then move into break out groups, each group engaging an attribute of the 4P heuristic. Each group will be presented with a "starter problem" and asked to consider it through the lens of their particular "P". Participants will be encouraged to move amongst the groups in order to become aware of each of the elements.

During the final 30 minutes, participants will convene to share strategies and insights along with areas of uncertainty and discomfort. Throughout all stages of the event, scribes will record ideas for sharing and further discussion after the event. The organizers and participants will envision next steps in the emerging research area of creative information practices and discuss the most suitable means for post-event follow-ups and ongoing communication.

---

1 Trauth has developed a creative, dramatic outlet as a means to “report” findings about her research on barriers imposed and internalized by women who aspire to careers in science and technology (Kiefer-Boyd, et al. forthcoming). See http://www.eileentrauth.com/theatre-of-social-engagement.html
Relevance to the Conference/Significance to the Field

This event speaks to the conference themes of innovation and wisdom. Scholarship-in-action, we suggest, involves tapping more deliberately into our creative capacities – which we seek to model throughout the session and related conference activities. The event seeks to appeal to a wide audience, from experienced researchers to new doctoral students - anyone with an interest or concern in creativity.

Other special features of the event

Observers will be on hand to “witness” activities throughout the event using tools based on Anderson’s ethnographic work at earlier conferences. In this way, the organisers seek to inform organizers of future iConference gatherings of the opportunities (realized and missed) of the event.

The event will also be making active use of the contemplative space at the iConference as a locale for enacting Anderson’s “Pause” in her 4P heuristic. Suggestions for this space include a period of silent sitting meditation or a life writing activity.

References


Acknowledgements

We thank Gary Burnett, Ameila Gibson, Anatoliy Gruzd, Michelle Kazmer, Bo Xie, and peer reviewers for their feedback and assistance.


Copyright is held by the authors.

Theory and the Social Nature of Information

Adam Worrall
College of Communication and Information
Florida State University
apw06@my.fsu.edu

Jes A. Koepfler
College of Information Studies
University of Maryland, College Park
koepfler@umd.edu

Jessica Vitak
College of Information Studies
University of Maryland, College Park
jvitak@umd.edu

Nicole D. Alemanne
College of Communication and Information
Florida State University
nalemanne@fsu.edu

Jaime Snyder
School of Information Studies
Syracuse University
jasnyd01@syr.edu

Howard Rosenbaum
School of Library and Information Science
Indiana University
hrosenba@indiana.edu

Abstract

Many scholars have discussed their use of theory and metatheory to study the social nature of information. The approaches brought to the table by doctoral students and junior faculty will continue to shape the future of the information field with a social perspective. This panel will present the theoretical frameworks used by four emerging information scholars (Koepfler, Vitak, Alemanne, and Snyder), who will describe the social nature of information in the context of existing information research and social theory. Each will answer three questions: How and why did they include particular theories in their research framework? How is their framework and its view of the social nature of information unique? What are the implications of their work for studying the social nature of information? The moderator (Worrall), panelists, and discussant (Rosenbaum) will encourage new thinking and discussion among researchers and practitioners interested in social and theoretically-based studies of information.

Keywords: Social theory, social nature of information, theory, framework, social perspective

Purpose

Applying a social lens to information is now common (Kling, 1999; Raber, 2003; Talja, Tuominen, & Savolainen, 2005), and extends back to the exchange and sharing of information Bush (1945) imagined for his memex. To study the social nature of information, theoretical frameworks apply different terms to the social contexts: groups, communities, networks, worlds, grounds, and others (Clarke & Star, 2008; Ellis, Oldridge, & Vasconcelos, 2004; Fisher, Durrance, & Hinton, 2004; Haythornthwaite, 2007; Jaeger & Burnett, 2010). No two frameworks are identical, but they share many similarities.

Scholars have discussed their use of theory and metatheory to study the social nature of information (Chatman, 2000; Fisher et al., 2004) or promoted a particular view (Hjørland, 1998; Tuominen & Savolainen, 1997). The approaches brought to the table by doctoral students and junior faculty will continue to shape the future of the information field with a social perspective. This panel will present the theoretical frameworks used by four emerging information scholars. They will describe the social nature of information in the context of existing information research and social theory, encouraging new thinking and discussion among an intended audience of researchers and practitioners interested in social and theoretically-based studies of information.
Activities

The panel will begin with the moderator, Adam Worrall, familiarizing the audience with the event structure and framing the social perspective of information. Each panelist will take ten minutes to present her area of research and the theoretical framework she is applying to it. Each panelist will address the following three questions:

- How and why did you include these theories in your research framework?
- How is your framework and its view of the social nature of information unique?
- What are the implications of your work for studying the social nature of information?

Following the presentations, a discussant, Howard Rosenbaum, will react to and reflect on the panelists’ frameworks and research. He will take ten minutes to offer a broader perspective on how the frameworks are similar and unique in the context of past, present, and future information research.

An in-depth, vibrant discussion involving the audience, panelists, and discussant will fill the remaining time, guided by the panel moderator. Potential topics include but are not limited to (a) similarities and differences between frameworks; (b) advantages and disadvantages of different frameworks and social theories; and (c) the application of frameworks and social theories to multiple research problems.

Our session will last 90 minutes. Attendees will network and collaborate with scholars who may differ in research problems of concern, but share interests in similar theories and metatheories. The session will also bring together scholars with shared interests in social and theoretically-based studies of information, promoting “valuable” new thought and consideration of social theories and their application to information research by the audience and panelists (Veinot & Williams, 2012, p. 11).

Panelists

Jes A. Koepfler is a doctoral candidate in the College of Information Studies at the University of Maryland. Her dissertation research focuses on identifying salient values expressed through informal communication (i.e. tweets) by stakeholders related to the issue of homelessness. Koepfler’s presentation will focus on the theoretical framing and methodological approaches she uses to identify salient values. She will discuss the practical implications this work has for better understanding information use and behavior from a values perspective among multiple stakeholders in an online context, drawing on theoretical frameworks from social psychology (Schwartz, 1992), information studies (Cheng & Fleischmann, 2010), and human-computer interaction (Friedman, 2011).

Jessica Vitak is an Assistant Professor at the University of Maryland’s iSchool, studying the social impacts of new communication technologies. Vitak will discuss how social capital—a sociological concept describing the resources individuals exchange with members of their social network—can be applied to studying communication technologies like Facebook. After providing a theoretical and methodological overview of bridging social capital (Burt, 1992; Lin, 2001; Putnam, 2000; Williams, 2006), she will discuss results from a recent lab study in which participants characterized the specific bridging resources they exchange through interactions with their Facebook “friends,” and how those resources relate to relational perceptions of network members.

Nicole D. Alemmane is a doctoral candidate in The Florida State University’s College of Communication and Information. Her dissertation research focuses on interdisciplinary academic teams as intrinsically transient social worlds. She will discuss her work in developing a theoretical framework with which to study knowledge co-creation processes in interdisciplinary teams whose work is time-limited. This research draws on previous investigations into social worlds and social processes, including the social worlds framework (Strauss, 1978; Clarke & Star, 2008), social capital (Coleman, 1988; Putnam, 2000), social networks (Haythornthwaite, 1996; Wellman, 1999), and group processes in intrinsically transient social worlds (Kazmer, 2006, 2010).

Jaime Snyder is a postdoctoral research fellow at the School of Information Studies, Syracuse University. Snyder’s research focuses on social aspects of visualization practices and the creation and use of visual information in social contexts. Her dissertation investigated spontaneous drawing during face-to-face conversations as an information-driven communication practice. Snyder’s work drew on theories from linguistic anthropology (Hanks, 1996) and interactional sociolinguistics (Gumperz, 1982) in order to expand the ways that images and image-making are studied in information science. Her
presentation will discuss notions of framing (Tannen & Wallat, 1993), footing (Goffman, 1979), and stance (Jaffe, 2009) that provided the analytic framework for this research.

Discussant

**Dr. Howard Rosenbaum** is the Associate Dean and an Associate Professor of Information Science in the School of Library and Information Science (SLIS) at Indiana University. He studies social informatics, ebusiness, and online communities; has published in a variety of information science journals; and co-authored the 2005 book "Information Technologies in Human Contexts: Learning from Organizational and Social Informatics" with Steve Sawyer and the late Rob Kling. Rosenbaum has also presented at ASIS&T, iConferences, and elsewhere. He has been recognized often for excellence in teaching and for the innovative use of technology in education, receiving awards and recognition from ASIS&T, Indiana University, Techpoint, and the Indiana Partnership for Statewide Education.

References


“It’s a Series of Tubes”:
Exploring Net Neutrality Policy through Critical Making

Jean-François Blanchette
Information Studies, UCLA
blanchette@ucla.edu

Matt Ratto
Faculty of Information, U. Toronto
matt.ratto@utoronto.ca

Abstract

As an increasing proportion of social life—from education to medicine, business, and national security—becomes mediated through the single medium of the Internet, what are appropriate policies to manage and fairly prioritize the flow of packet traffic through networks? Our critical making exercise will use a car track set as a way to experiment with network management policies and the value choices they imply.

Keywords: critical making, network neutrality, internet policy, materiality of computation.

As an increasing proportion of social life—from education to medicine, business, and national security—becomes mediated through the single medium of the Internet, what are appropriate policies to manage and fairly prioritize the flow of packet traffic through networks? In recent years, the issue has become polarized in terms of a “Net Neutrality” debate that pits two camps against one another. On the one hand, activists and scholars such as Larry Lessig and Tim Wu argue that an essential ingredient of the Internet’s power to foster innovation lies in the non-discrimination design principle that treats all packets equally, regardless of their origin, destination, or purpose. On the other hand, network service providers argue that, given the limitations and high costs of communication infrastructures, they must resort to network management practices in order to maximize service for all users. This is particularly the case given constant growth in network traffic due to new content distribution platforms (Netflix videos account now account for half of all internet traffic), higher resolutions devices (e.g., Apple’s “Retina” displays), or new applications (e.g., Siri). Network providers use both technical and economic approaches to network management: technical approaches include throttling, caching, and optimization, such as transcoding of video and images; economic approaches include tiered service (e.g., limited vs. unlimited data plans), peak/off-peak pricing, and bundling (e.g., family plans). The FCC has currently given free reign to wireless service providers to implement any “reasonable” management practices they deem necessary, as long as they “transparently” disclose them (see http://www.fcc.gov/topic/open-internet).

This workshop proposes to explore network management practices and the politics of net neutrality through a critical making experiment. Critical Making, a mode of engagement that links critical reflection and physical production, supplements previous forms of policy analysis through the creation of specific, goal-oriented experiments. These experiments are intended to re-map the relationship between conceptual and linguistic forms of social analysis (e.g. social scholarship), and materially and embodied forms of shared making. In past work (Ratto & Hockema, 2009; Ratto, 2011a; Ratto, 2011b ) we have found these techniques to be useful in exploring stakeholder’s critical issues around new technological developments. Critical Making experiments have been carried out in London, Amsterdam, Scotland, the US, and in Canada. (see previous citations for specific examples.)

The planned event will explore issues of net neutrality as well as possibilities, potentials, and challenges of using critical making as a tool for policy support and academic reflection. Our focus for this workshop will be two-fold; first to use a focused critical making exercise to explore questions relative to the allocation of scarce computational resources; and second, to review our shared process from a more ‘meta’ level and to think about how it supplements and extends more traditional types of policy analysis.


Copyright is held by the authors.
The problem of maximizing packet traffic on networks is largely identical to that faced by engineering and policy makers as they seek to maximize road traffic on the congested roads of cities. In this case as well, economic (e.g., toll lanes) and technical (e.g., metered highway entrances) approaches are both used to maximize service and prioritize certain types of traffic. Our critical making exercise will use a car track set as a way to experiment with network management policies and the value choices they imply. Working in teams, session participants will use software and hardware pre-assembled by the conference organizers to build simple radio-controlled robot vehicles. These vehicles will represent individual packets on a network. Participants will customize both a central controller and the physical landscape in order to explore mechanisms for differential packet management. We will then use the building and testing process to think and talk about the above topics, and about how shared ‘making’ practices can supplement traditional qualitative and quantitative policy analysis.

Participants will be required to bring laptops (one per team) to the session, be prepared to install the Arduino software IDE, and to engage in hands-on work. All materials will be provided by the organizers. **NO EXISTING TECHNICAL EXPERTISE IS REQUIRED.** If possible, the resulting objects could be displayed during the conference at the poster session or other event.

**Workshop Plan**

The workshop will require about 3 hours and will consist of 5 phases (we thus require 2 back-to-back sessions of 90 minutes):

1. A brief explanation and exploration of critical making, of the net neutrality debate, and of current network management practices. This will include introductory and framing remarks by the organizers and group discussion.
2. An overview of the technologies being used, guided installation of necessary software, and description of the pre-assembled toolkits.
3. A hands-on assignment involving designing, building, trouble-shooting, and testing the prototypes.
4. Group discussion of the projects, their commonalities and differences, what they offer for critical reflection on resource sharing and allocation.
5. A final discussion on critical making as a tool for exploring policy issues.

All materials will be provided by the organizers. The workshop should be limited to 25 participants. A/V requirements are a Mac-compatible video projector. The room should be reconfigurable with movable tables and chairs.

The workshop will build on similar events and workshops held at previous iConferences, including the 2012 “Values|Making|Critique|Design” workshop, the 2012 “Design Jam in iSchools’ even, the 2011 “Storytelling, Narratives and Metaphors in the Design and Use of ICTs” workshop, and the 2011 Poster and Roundtable session, “Design Methods for the Information School Curriculum.” These workshops and events have shown that iConference attendees share a significant interest in design issues and methodologies. We aim to attract such participants as well as those interested in internet policy. The workshop’s design and aim are central to the mission of the ISchool movement, to connect technology, information, and people. It makes uses of innovative pedagogical techniques that fuse together technical and sociological exploration.
Design Jams in iSchools: Approaches, Challenges and Examples

Michael B. Twidale
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
twidale@illinois.edu

Ingbert Schmidt
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
ifloyd2@illinois.edu

Jeff Ginger
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
ginger@illinois.edu

Peter Organisciak
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
organis2@illinois.edu

Piotr Adamczyk
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
pdadamcz@illinois.edu

Christopher Lueg
School of Computing and Information Systems
University of Tasmania
Christopher.Lueg@utas.edu.au

Abstract

Through a live demonstration, we will showcase a group of focused design techniques known collectively as a Design Jam. Design jams are about looking at a particular design challenge and thinking-by-doing. Although they often have a component of brainstorming, they involve additional activities, including paper prototyping, and storytelling with personas and scenarios. After the design jam, we will share experiences of teaching design techniques in ischools.

Keywords: design thinking, teaching design, and paper prototyping

Alternative Event Description

Through a live demonstration, we will showcase a group of focused design techniques known collectively as a Design Jam. Activities that fit this broad definition are sometimes also called charettes, sprints, hackfests and barcamps. Design jams are about looking at a particular design challenge and thinking-by-doing. Although they often have a component of brainstorming, they involve additional activities, including paper prototyping, and storytelling with personas and scenarios.

We want to show the power of thinking-by-doing-design as a technique, one that has great potential in an iSchool setting even with people who are unfamiliar with it. This event will build on the success of the Design Jam alternative event offered at the 2012 iConference, updated based on the comments and reactions of last year’s participants.

Intended Audience

There are two kinds of participant who will benefit from this event:

- Those who teach or plan to teach design in iSchools.
- Those who are unfamiliar with the approach and curious, but perhaps find it to be an alien way of thinking, or are skeptical of what its proponents claim for it and want to see it working in action.


Copyright is held by the authors.
Proposed Activities

Invitation to Participate

Prior to the event, we will create and publicize an invitation to participate. This will be a website (https://sites.google.com/site/iconf13design/) describing the idea of the session, some design challenges, and giving an opportunity for people to indicate their intention to participate. Our design challenges are meant to give a foundation for thinking of the design jam as a method in the context of actual problems. One theme this year is designing for privacy management: ways to make it easier for people to understand what they are sharing, with whom and what the consequences are.

Explanation for Observers

We will create a flyer to hand out to people arriving to the event after it has begun so they can understand what the design teams are doing, and encouraging them to observe the process.

Design Jam as Public Spectacle

For the first hour of the event, participants will be in teams working on one of a number of different design challenges. The aim is not to have a competition, but to playfully explore design spaces and see how far an idea can get in a short period of time. Other attendees will be welcome to come and watch the design activities as they unfold. A number of interpreters (see below) will be on hand to explain the process to those unfamiliar with the idea of a design jam.

Presentation of Results

Teams will show what they produced; briefly noting what was unexpectedly easy and fast, or slow and difficult.

Discussion

The rest of the event will be a plenary discussion of Design Jams and similar approaches in iSchool settings, with a focus on teaching design skills and design thinking. Possible topics are:

- How to get started with design jams?
- Variants to the design jam idea?
- What makes a good design challenge?
- How can design jams go wrong and how can you address that?
- How to nurture a maker culture in an iSchool?
- How to fold iterative prototyping into community informatics / bioinformatics / IR?
- How iterated design jam experiences can build up both skills and concepts.

We invite participants to share their experiences (positive and negative) of using various design activities in teaching. We find that having had an experience of participating in a design jam prior to a discussion, all participants have a ready to hand set of examples to explore issues of pedagogy, approach, relevance and concern. It means that those who are unfamiliar with the idea of design jams can participate actively, by sharing their first impressions.

Follow-through

We will add to the pre-conference website to share readings, design challenges, pedagogies and resources.
Facilitation and Interpretation

We have a number of co-organizers and colleagues who plan to attend and will serve as facilitators to design teams. The organizer will explain what is going on to observers, people arriving late and those unused to design jams, design thinking or design at all.

Relevance to the Conference/Significance to the Field

We believe that user-centred design involving different kinds of participatory inclusion has a natural home in iSchools, in research, teaching, and service.

Although there is already design work based in iSchools, we believe that there is a potential for much more. The multidisciplinary ethos of iSchools and the focus on the interactions between people, information, and technology create a great opportunity to design better, more effective applications and indeed socio-technical systems. Drawing on the described purpose of the iConference on its website, we think that this is a classic case of an opportunity to advance the boundaries of information studies.

Design Jams also have the potential to help multidisciplinary teams to share ideas and insights. A technological focus on using computational tools can limit the comfort of those who do not think of themselves as ‘techy’, while a more analytic conceptual analysis of a problem can alienate those who lack familiarity with the frameworks and terminology deployed. Both of these approaches can lose people through the various layers of abstraction they necessarily deploy. However, many people feel comfortable grabbing a crayon and drawing ideas of what they would like the application to do, and then iterating on that with others.
Igniting Talk on Digital Literacy

Ingrid Erickson  
Rutgers University  
ingrid.erickson@rutgers.edu

Eric Meyers  
University of British Columbia  
eric.meyers@ubc.ca

Ruth Small  
Syracuse University  
drruth@syr.edu

Abstract

New technologies and developments in media are transforming the way that individuals, groups and societies communicate, learn, work and govern. This new socio-technical reality requires participants to possess not only skills and abilities related to the use of technological tools, but also knowledge regarding the norms and practices of appropriate usage. To be ‘digitally literate’ in this way encompasses issues of cognitive authority, safety and privacy, creative, ethical, and responsible use and reuse of digital media, among other topics (Meyers, 2009; Arnone, et al., 2011). A lack of digital literacy increasingly implicates one’s full potential of being a competent student, an empowered employee, or an engaged citizen. Digital literacy is often considered a school-based competency, but it is introduced and developed in informal learning contexts such as libraries, museums, social groups, affinity spaces online, not to mention the home environment. This community-building event will recognize and connect the ways and places we conceptualize and realize digital literacy.

Keywords: digital literacy, informal learning, socio-technical, digital media

Introduction

This series of Ignite Talks will provide a platform for members of the iSchool community to develop a deeper understanding of digital literacy and its importance within an iSchool research agenda. Via talks and discussion, we will showcase a few of the many ways that digital literacy affects the lives of people as they explore media representations, solve everyday problems, and create new knowledge.

The intended audience for this forum is scholars at all career stages who have an interest in conceptualizations of digital literacy, pedagogies and practices for fostering digital literacy, or the policy implications of digital literacy in society.

Session Overview

We have broken the 90-minute alternative event session into two parts. During the first part, following a brief introduction, Ignite Talk speakers will present for 5-minutes using up to 25 slides. These presentations will be followed by 5 minutes of Q&A. The second part of the session is meant to be synthetic and interactive. We will invite the audience to discuss the ignite talks in small groups, clustered around the individual presenters who will lead the discussion and summarize the group response. The outcomes of each small group discussion will be displayed using sticky notes and white boards. The final 10 minutes of the discussion will involve participants’ free movement among the groups to make their own sense of the discussions. The schedule for the session is as follows:

- [5 mins] Introduction/Welcome
- [10 mins] Ignite Talk 1 + 5 mins Q & A
- [10 mins] Ignite Talk 2 + 5 mins Q & A
- [10 mins] Ignite Talk 3 + 5 mins Q & A
- [10 mins] Ignite Talk 4 + 5 mins Q & A
- [10 mins] Ignite Talk 5 + 5 mins Q & A
- [35 mins] Facilitated Interaction

Copyright is held by the authors.
Invited Speakers and Audience Participation

As indicated above, our current schedule allows for five Ignite presentations. A set of potential speakers has been invited, but cannot be confirmed at the time of this publication. We are motivated to curate a group of recognized and emerging scholars active in the domain of digital literacy to provide a diversity of insights to understanding the topic—particularly as this approach best befits the iSchool community. We welcome any number of participants to join as members of the audience and follow-up discussion.

- Katie Davis (University of Washington)
- Ingrid Erickson (Rutgers University)
- Sara Grimes (University of Toronto)
- Marcia Mardis (Florida State University)
- Paul Marty (Florida State University)
- Eric Meyers (University of British Columbia)
- Rebecca Reynolds (Rutgers University)
- Sarita Yardi Schoenebeck (University of Michigan)

Relevance to the Field

We anticipate that this alternative event will yield two outcomes of benefit to the field. First, we will build on the participation and feedback of the attendees to organize a workshop or research paper roundtable on digital literacy for the 2014 iConference. Additionally, we will use the insights from the discussion to write a summarizing article post-conference about the place of digital literacy in iSchool discourse for either Interactions or First Monday. We expect each of these outcomes to encourage a deeper attention to issues of digital literacy within future iSchool scholarship.

References


Tech Learning Potluck

Michael B. Twidale  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
twidale@illinois.edu

Sally Jo Cunningham  
Department of Computer Science  
University of Waikato  
sallyjo@cs.waikato.ac.nz

James Howison  
Information School  
University of Texas, Austin  
james@howison.name

Peter Organisciak  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
organis2@illinois.edu

Jeff Ginger  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
ginger@illinois.edu

Ingbert Schmidt  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
ifloyd2@illinois.edu

Brittany Smith  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
bnsmith3@illinois.edu

Karen Baker  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
Ksbaker2@illinois.edu

Kate Williams  
Graduate School of Library and Information Science  
University of Illinois at Urbana-Champaign  
katewill@illinois.edu

Abstract

We will share experiences of learning and teaching technology, exploring pedagogies, barriers to learning, and recurrent misconceptions that arise in learning to use computer applications. To ground the discussion, we will start the event by each taking turns in both learning and teaching how to do something with a particular application.

Keywords: Learning Technology, Teaching Technology, Storytelling, Pedagogy.

Overview

We are all subjected to an onslaught of new gadgets and applications for our computers, tablets and phones. New applications and new versions of old applications are percolating into our work lives and our family lives, our health and exercise activities and our social interactions, even into religious practices.

It is most unlikely that a training course will be on offer at the time we need it. Many people cope by learning informally from friends, family and colleagues, or ‘playing around’ or fiddling with the application. This is often successful, but a lot of people find this way of learning technology rather alien or daunting, and even technophiles have had frustrating learning experiences when things went wrong. It is all the more galling when the application’s marketing proclaims that it is “easy to use”, “intuitive” or “so simple a child could use it”. Failures in the light of such proclamations can make you feel old, inadequate and obsolete.

But does it need to be that way? Are some people wired to be good at learning technologies, while others remain bewildered? Or perhaps it is simply that some people have some skills and a degree of confidence that others lack. Can we teach the skills that will help people be more confident and effective in learning technologies? Are iSchools the place to take ownership of this challenge? The topic is a wonderful mixture of research, teaching, and service that plays to many of our strengths and interests.
Alternative Event Description

We will explore these ideas by sharing our experiences—both successful and frustrating—of learning technologies ourselves and helping others to learn themselves.

Intended Audience

People who are involved in teaching technology, especially those interested in enabling learners to take greater control of their technological learning and coping as a lifelong endeavor. This is likely to include but not be limited to those interested in the learning of technology by children, undergraduates, graduate students, office workers, scientists, older people, people with various disabilities, and socially disadvantaged or marginalized groups. Although much technology learning takes place outside formal classroom settings, we want to invite participation from those who teach technology classes, those who study or facilitate more informal or incidental technology learning, and those who observe technological learning, helping, adoption or appropriation in a particular setting.

First Activity: Sharing Skills in a Tech Learning Potluck

Be prepared to teach someone how to do something with a computer application in five minutes. Pick an application that you like using, perhaps one that improves your productivity or your life in some small way. Your chosen application could run on a laptop, smartphone, or tablet computer. Ideally, you will bring along the device with you (we will have wifi access). But if you do not have a device, bring a skill that you can teach on someone else’s networked device. A web-based service might be a good idea. For example maybe you know how to do something clever with Google calendar that you suspect that others do not. If you have a number of small things you can teach, that’s even better. We will form up into pairs (facilitated by the organizers) with each participant trying to teach another how to use their application to do something, and learning a new one in turn. We will then repeat with a different pairing, and if time permits, a third.

Second Activity: Sharing Stories about Tech Learning and How to Support it

Inspired by these immediate experiences of learning and helping others to learn a technological feature, we will discuss aspects of the process of tech learning, including various barriers and difficulties and how to address them. Participants will be encouraged to illustrate issues by examples from the potluck as well as from their wider experience. By having a rich set of examples of challenges and approaches we will be in a better position to observe both recurrent themes (similar things in different settings) and variations (different versions or outcomes of a similar issue). Storytelling is a particularly effective way of sharing ideas and issues around contextualized learning of technology.

We will talk about various barriers to tech learning and the strategies and tactics people have used to facilitate or demystify it. Sometimes the challenge is just understanding what needs to be learned, the prerequisites and what to learn first. Often the issue is the larger need to learn how to learn. Some likely topics include:

- Do we know how to build “technical self efficacy”?
- Distorted perceptions of the costs, risks and benefits of tech learning.
- Understanding different kinds of technology or computer phobia.
- Issues in “learned helplessness” from bad tech experiences.
- Classic recurrent misconceptions about computational technologies, learning, and tech learning.
- Approaches, strategies and tactics that participants have used to address the above.
- Ways to facilitate tech learning, including productive interactions with online communities.
- Learning skills and techniques that people may be lacking.
- Ways to overcome gender biases in tech fiddling.
- Supporting people in learning to do what they want to do, not what we think that they should do.
How is it different if the people doing the tech learning are children, adults, graduate students, professors, office workers, elderly people, someone trying to apply for an entry level job using their first ever online application form in a public library, etc.?

Which issues recur regardless of who is learning, what or where?

But no doubt others will emerge out of our discussions. We have created a Google Doc for participants to share stories, techniques, references and resources: http://goo.gl/7U60p

This will be used in preparation for the event, as a way of sharing what emerges during the event, and as a subsequent resource.
Collaboration in Action: Enabling Innovative Scholarship with Social and Crowdsourcing Services

Chirag Shah
Rutgers University
chirags@rutgers.edu

Daqing He
University of Pittsburgh
Dah44@pitt.edu

Marie L. Radford
Rutgers University
mradford@rutgers.edu

Jung Sun Oh
University of Pittsburgh
jsoh@pitt.edu

Lynn Silipigni Connaway
OCLC
connawal@oclc.org

Abstract

With an exponentially growing set of e-services and social networks that allow people to be not only consumers, but also producers of information, information seeking and sharing behaviors are rapidly changing. Innovations in areas such as information exchange and knowledge management are coming from scholarship in data sciences, and the "wisdom of the crowd" has become more than a passing trend. The focus of this event would be to discuss the latest developments in the field of social media and crowdsourcing specific to information seeking, knowledge management, and innovative methods for collaborative question-answering. Specifically, the event will facilitate discussions about and engage the audience in topics such as social search, community-based question-answering, and hybrid models for information seeking. These discussions will be guided by the organizers who come from a variety of backgrounds, institutions, and research areas.

Keywords: information seeking behavior, crowdsourcing, social media, social Q&A (SQA), virtual reference (VR)

Purpose and Intended Audience

This event is designed for an audience of researchers, faculty, Ph.D. students, practitioners, and members of the information industry. It is intended to bring researchers in the fields of social and collaborative information seeking, crowdsourcing, social media, as well as virtual reference (VR) and Social Question & Answer (SQA) services together to discuss their recent research findings and to set a course and agenda for future work. Thus, primary objectives are to:

- Share our expertise and research findings
- Discuss limitations and possibilities for future synergies
- Identify ways to share data and other resources
- Suggest and plan possible collaborations

Particularly, the discussion during this event will be approached from different angles - How does community participation affect information-seeking behaviors? What happens when search goes social? What could social search and VR services learn from each other to better serve users?

Acknowledgements: Some of the work reported here is being conducted as part of the research project Cyber Synergy: Seeking Sustainability through Collaboration between Virtual Reference and Social Q&A Sites, funded by the Institute of Museum and Library Services (IMLS), Rutgers, The State University of New Jersey, and OCLC, Inc. (http://www.oclc.org/research/activities/synergy.html). A portion of the work presented here is also supported in part by Wuhan International Science and Technology Cooperation Fund under agreement 201070934337 and National Science Foundation of USA under the agreement NSF/IIS 1052773.


Copyright is held by the authors.
Event Agenda, Ramp-up (development), and Follow-through

Unlike many traditional panels, this event is intended to be highly interactive, with a clear emphasis on audience participation. For instance, after making a few short presentations (30-45 minutes), the organizers will present specific topics and/or research questions to initiate conversations with the attendees. Examples include:

- How are information-seeking and sharing behaviors changing and shaping the new landscape of information sciences and practices? Think about examples and narratives.
- What is social search? How does it inform “traditional” searching behaviors and models, and what challenges and opportunities may be ahead?
- How do we compare expert-based (e.g., VR) with crowd-based (e.g., SQA) services? What ideas, components and practices can each learn from one another?

Finally, the organizers will encourage discussions that lead to planning for future research and potential collaborations.

Before the event (and the conference), the organizers will set up a website to facilitate organization of themes and research questions among the organizers. This website also will be open to others, serving as an important tool to ramp-up the pre-event discussions.

During and after the event, the organizers will setup a blog and/or a wiki to facilitate discussion and planning. The organizers also will create a hash-tag on Twitter for participants (both on and off site) to share their comments and engage in discussions easily.

After the event, the blog and/or the wiki pages will be kept alive, allowing researchers to share their data and findings, and continue connecting with each other for discussions and collaborations.

Relevance to the Conference and Significance to the Field

The focus of this event fits nicely with the conference’s theme “Scholarship in Action: Data, Innovation, Wisdom.” Given both the vast research opportunities and the nontrivial challenges the large-scale social data present to the iSchool community, we believe that it is critical for researchers and practitioners with shared interests to learn from one another, share innovative approaches, and eventually develop collective wisdom. This event will serve as an initial step towards this goal. The research findings and discussions shared through the event and the post-event activities will encourage exploration for the creation of new technological and conceptual configurations, such as possibilities for collaboration and data exchange of social, crowdsourcing, and one-on-one VRS. As the majority of the audience conducts research or engages in practice in information sciences or services, they will find the focus relevant and informative with regard to their academic interests and responsibilities.

The format of this event is highly social and interactive. It provides first-hand experience for both the organizers and audience to work on social data, on information sharing, and on community building and collaboration.
Methods, Practice, Theory: Researching Human Information Behavior in the Context of Migration

Jessica Lingel  
Rutgers University  
jlingel@eden.rutgers.edu

Lisa Quirke  
University of Toronto  
Lisa.quirke@utoronto.ca

Nadia Caidi  
University of Toronto  
nadia.caidi@utoronto.ca

Karen Fisher  
University of Washington  
fisher@uw.edu

Ellen Pozzi  
William Patterson University  
pozzie@wpunj.edu

Abstract

This session uses a roundtable format to bring together researchers working on human information behavior (HIB) and non-dominant groups. As a session, we focus on the study of transnational migration, with the understanding that many of the issues faced by researchers in this area are shared by projects on other non-dominant groups. Our objective is to foster conversation among a community of active researchers on the theoretical, methodological, professional and political aspects of migration. The term migration is here considered broadly, so as to include not only movement between countries, but also movement between the urban and rural, permanently or temporarily, once or repeatedly. Rather than discussing individual research projects in depth, we will share methodological challenges of studying migration; relevant interdisciplinary theory or theories that provide insight into migrational experience in the context of information practices; and extensions of research to professional work of information institutions. Our key objective is to provide a space that brings together a community of researchers who share similar interests in (and face similar challenges with) qualitative research with non-dominant groups.

Keywords: immigration, methodology, theory, qualitative research

Introduction

In this session, we will explore key methodological, practical and theoretical issues in information behavior research related to migration, bringing together themes shared by research to date and pointing to possibilities for future scholarship. Although our focus is on HIB research related to migration, themes and concepts developed in this discussion bear relevance to scholarship in other areas of information science, including work on marginalized groups and subcultures, and outreach programming at public libraries. We have three key conceptual foci: theory, methods and practice.

Theory

Our session works towards gathering research in this area into a more cohesive collection of existing LIS research on migration, as well as making connections to outside theory. At the same time, we question the extent to which it is useful to try and build a cohesive body of scholarship around migration. Are migratory experiences so heterogeneous as to defy cohesion, at least on some level? In the midst of fuzzy conceptions of different, yet often profoundly distinct, populations, can findings from studies on migrant workers and international students be applied to refugees or immigrants? What about differences between domestic minorities and those who have experienced migration? On a disciplinary level, we also probe the tendency to draw on theory from other disciplines (including sociology, immigration studies,
urban studies, science and technology studies, cultural studies, economics) rather than building theory from our own discipline. In doing so, we encourage LIS researchers not just to poach but to produce theory.

Methods

To explore fully the contexts of information behavior and identify relevant variables that shape it, qualitative, exploratory studies are needed. This panel focuses on the range of qualitative and interpretive methods for studying migration, drawing on research experience to discuss affordances and limitations of different methodological decisions. Although we focus on methods that have been used for research on migration, these methods lend themselves to qualitative work in a wide array of areas.

Practice

Though information behavior research in the context of migration is important to the work of information professionals, there are many unexplored areas and topics that need to be resolved for future studies to be able to provide more concrete findings to practitioners. These include: a consideration of the local policy context of migration; an understanding of the context within which individuals are migrating; and an exploration of the types of challenges individuals face during or after migration. By understanding the types of challenges and barriers individuals face during and following migration, research will be able to show which barriers exist and can be solved at the individual level, and which are systemic issues that require resources offered at a broader scale, such as public education campaigns.

Contributions

Current migration-focused research in information studies is scattered and fragmented, and there is furthermore a lack of dialog across disciplines. Some concepts, for instance information gatekeepers (see Barzilai-Nahon, 2009), haven't gained traction among others studying immigrant information behavior. At the same time, although quantitative, survey-based studies of immigration are plentiful, there are few empirical studies explicitly of immigrant information behavior; of those that exist, some are large-scale, descriptive and survey-based (i.e. Mehra & Papajohn, 2007) which neglect many of the micro level factors – language, country of origin, immigration class, family status, social context – that can greatly shape information behavior; others focus on institutions such as libraries and other organizations and their role in immigrant adaptation without fully exploring immigrant behavior independent of these programs or facilities (e.g. Burke, 2008). While there is a large LIS practitioner-based literature, these articles are typically brief descriptions of library programs for immigrants; although helpful in understanding the successes of specific libraries in serving certain populations, these texts do not provide nuanced analysis of immigrant information behavior. By bringing together scholars with different areas of expertise on migrational information behavior, different methodological approaches and different theoretical frameworks, this session will spark dialogue not only in terms of theory, but also in terms of methodology and professional practice.

Format

Instead of a traditional panel where participants present individual work, organizers will briefly introduce themselves and their research. They will then focus in turn on theory, methods and practice, with the objective of building a rich conversation about migration and HIB. Dr. Karen Fisher will act as a respondent, drawing together themes from other organizers and soliciting audience participation. Our goal is to foster conversation and create a space of skill sharing. Dialogue between panelists, and between panelists and the audience, is vital to developing the most complex understanding of needs for LIS research in this area.
References


Help iDigBio Reveal Hidden Data: 
iDigBio Augmenting OCR Working Group Needs You – Part II

Deborah Paul  
Florida State University/iDigInfo, iDigBio  
dpaul@fsu.edu

P. Bryan Heidorn  
University of Arizona/School of Information and Library Science  
heidorn@email.arizona.edu

Edward Paul  
Florida State University/iDigInfo, iDigBio  
dpaul@fsu.edu

Jason Best  
Botanical Research Institute of Texas (BRIT)/Biodiversity Informatics  
jbest@brit.org

Amanda Neill  
Botanical Research Institute of Texas (BRIT)/BRIT Herbarium  
anneill@brit.org

John Mignault  
New York Botanical Garden  
jmignault@nybg.org

Abstract

Integrated Digitized Biocollections (iDigBio) is a nation-wide effort funded by the National Science Foundation (NSF) to digitize data from hundreds of millions of natural history museum specimens. In a concerted five-part outreach effort, the iDigBio Augmenting Optical Character Recognition Working Group (A-OCR WG) coordinated a 2013 iConference Workshop, Poster, Notes submission, Alternative Event and a concurrent Hackathon hosted by the Botanical Research Institute of Texas (BRIT). The Workshop titled, "Help iDigBio Reveal Hidden Data: iDigBio Augmenting OCR Working Group Needs You" introduces the iSchools community to iDigBio and the A-OCR WG mission and challenges to improve digitization efficiency. This related Alternative Event provides the A-OCR WG an opportunity to report back to iConference Workshop attendees about our first experience using a Hackathon model to work on parsing and user interface design issues specific to our needs. We anticipate to a lively, open discussion with event attendees and future collaborators.

Keywords: OCR, natural language, information analysis, machine language, iDigBio hackathon

Report Back from the iDigBio A-OCR Hackathon

In this iConference 2013 Alternative Event, "Help iDigBio Reveal Hidden Data: iDigBio Augmenting OCR Working Group Needs You – Part II" several of the A-OCR WG members return to update the iConference attendees on the results of the A-OCR Hackathon at BRIT. The hackathon (Table 3) focuses on parsing of OCR output and the user interfaces for human-in-the-loop interaction with the OCR output and digitization workflow process. We look forward to those attending the iConference iDigBio A-OCR Workshop on Tuesday coming back on Friday (Table 4) to hear about our experience and discuss the results. The informed conversation will help shape the future direction of our working group.

Overview. iDigBio is funded under the NSF’s Advancing Digitization of Biodiversity Collections (ADBC) program (https://www.idigbio.org/content/nsf-adbc-program-information) to build a cyberinfrastructure to integrate and expose museum specimen data for the public and researchers alike. In addition, iDigBio also has a directive to facilitate the development and dissemination of effective practices in digitization across museum collection types. Our diverse A-OCR WG is targeting improvement of extant workflows using OCR strategies as one way to speed up databasing throughput and assure data is fit-for-research-use.

Workshop. In our Tuesday afternoon iConference 2013 workshop, Help iDigBio Reveal Hidden Data: iDigBio Augmenting OCR Working Group Needs You, we introduce ourselves, our goals and our challenges (Table 1, Table 2). The A-OCR WG represents members of the natural history museum community who are actively engaged in digitizing specimen data from typed, hand-written and printed
labels from millions and millions of specimens currently only accessible to researchers with direct access to the physical specimen.

Table 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Title of talk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deborah Paul</td>
<td>Introducing iDigBio and the Augmenting OCR Working Group</td>
</tr>
<tr>
<td>Amanda Neill</td>
<td>Digitization of biocollections: a grand challenge in scope, scale, and significance</td>
</tr>
<tr>
<td>Jason Best</td>
<td>The Apiary Project: a workflow for text extraction and parsing for herbarium specimens</td>
</tr>
<tr>
<td>Edward Gilbert</td>
<td>Symbiota: Creating an OCR and NLP enabled user interface and workflow to efficiently digitize 2.3 million lichen and bryophyte specimens</td>
</tr>
<tr>
<td>Bryan Heidorn</td>
<td>HERBIS/LABELX: Machine Learning Approach to Parsing OCR Text</td>
</tr>
<tr>
<td>John Mignault</td>
<td>Linking Data: Biodiversity Heritage Library: supporting knowledge discovery from digitized content</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>130 – 145</td>
<td>iDigBio and the Augmenting OCR Working Group</td>
</tr>
<tr>
<td>145 – 315</td>
<td>Panelists’ Talks, each 15 – 20 minutes</td>
</tr>
<tr>
<td>315 – 500</td>
<td>Breakout Groups – Your Innovation and Wisdom</td>
</tr>
<tr>
<td>500 – 530</td>
<td>Wrap-up: Report Backs from the Breakout Groups</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Time</th>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>900 - 500</td>
<td>Wed., Feb. 13th</td>
<td>First Hackathon of the iDigBio Augmenting OCR Working Group hosted by the Botanical Research Institute of Texas (BRIT)</td>
</tr>
<tr>
<td>900 - 500</td>
<td>Thurs., Feb. 14th</td>
<td></td>
</tr>
</tbody>
</table>

Our iConference Workshop panelists represent a broad perspective on the topic of digitization in general and unique views on how OCR and use of OCR output fits in our overall mission. From biologists to programmers, information scientists and educators, our panelists’ vantage points may intrigue those in Information and Computer Science to think of new ways to apply their unique collective knowledge to our tasks.

Hackathon. In our early efforts to pick topics to work on, parsing and user interfaces stand out. In addition to iConference participation, we designed a hackathon with three specimen-image label sets varying in difficulty from simpler images with mostly printed or typed text to those images with a lot of handwriting or other miscellaneous objects present. Our hackathon participants are members of the working group and invitees selected from an open application process reaching outside our community’s borders.

Programmers, working with scripters and end-users, will utilize data dictionaries and authority files for the cleanest OCR output possible and then attempt to parse the resulting output into standard fields selected mainly from the Darwin Core (Wieczorek et al., 2012) data standard. We created a gold standard set of parsed records for comparison and scoring. For more details about our hackathon, please read our Call for Participation (http://tinyurl.com/aocrHack).

Because a human-in-the-loop appears to be integral to the most efficient use of OCR, OCR output and effective parsing, then discussion surrounding the development of better user interfaces is anticipated to be an important conversation at our hackathon, the iConference Workshop and at this Alternative Event. More than one-fourth of the hackathon participants develop user interfaces designed for manipulating OCR output.
**Alternative Event.** On Friday, this Alternative Event (Table 4) gives the A-OCR WG a 45-minute forum to tie all the outreach activities together. We will have time to share, with the iConference Workshop attendees from Tuesday and others, key observations and experiences from our first Hackathon and get input for development of our next target tasks.

Table 4

<table>
<thead>
<tr>
<th>Time</th>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>tba (45</td>
<td>Fri., Feb. 15th</td>
<td>iDigBio and the Augmenting OCR Working Group: report back after the hackathon</td>
</tr>
<tr>
<td>minutes)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The content is timely, that of getting useful research data to scientists and the public that is otherwise very difficult to access at all. Data included in our efforts will be used by scientists to address such issues as climate change, global warming, species discovery, niche modeling and host-insect-parasite relationships useful for agriculture and economics research. The data itself is to be integrated into a searchable cyberinfrastructure allowing scientists to create unique data sets for study.

Through iDigBio Workshops, Wiki, Working Groups, materials development, education and outreach, everyone is able to learn from and encouraged to contribute to our collective knowledge on OCR, improving algorithms for parsing OCR output effectively to populate databases, and user interfaces for these activities.

**Conclusion**

We are striving for new collaborations with those in Information Science and Computer Science who have expertise and interests in such areas as: information retrieval, information services, research methods and qualitative data analysis. Our group recognizes there are experts in the broader community whose participation and contribution would add greatly to our chances of success at addressing the issues on our Wish List ([http://tinyurl.com/OCRHackathonWishList](http://tinyurl.com/OCRHackathonWishList)) including improving outcomes using: image segmentation, autocorrection of typographical errors, normalization of data and such natural language processing tasks as automated text segmentation and markup. We look forward to connecting the natural history collections digitization community to the iSchools community through this iConference Alternative Event and related activities.

**Reference**


Acknowledgements: iDigBio is kindly funded by a grant from the National Science Foundation's Advancing Digitization of Biological Collections Program (#EF1115210). To each and every member of the Augmenting Optical Character Recognition working group, many kind thanks to you: Robert Anglin - Symbiota Developer; Jason Best - Biodiversity Informatics Director (BRIT); Renato Figueiredo - iDigBio IT; Edward Gilbert - Developer of Symbiota, Programmer; Nathan Gnanasambandam - Senior Research Scientist, Xerox; Stephen Gottschalk - Curatorial Assistant, New York Botanical Garden; Elspeth Haston - Assistant Curator: Digitisation, Royal Botanic Garden Edinburgh; P. Bryan Headorn - Director School of Information and Library Science, University of Arizona; Daryl Lafferty - SALIX Developer; Peter Lang - Pre-Sales Representative, ABBYY USA; Gil Nelson - iDigBio Digitization; Deborah Paul - iDigBio User Services; Anna Saltmarsh - Digital Collections Manager, Kew Herbarium; Nahil Sobh - InvertNet Developer, Researcher, University of Illinois at Urbana-Champaign; William Ulate - Technical Director, Biodiversity Heritage Library; Kimberly Watson - Curatorial Assistant, New York Botanical Garden; and Qianjin Zhang - Masters Student, University of Arizona. We must also thank Reed Beaman, Peter Obozinski, Elvis Hsin-Hui Wu, Karl-Heinz Steinke, Dean Pentcheff, Matt Yoder, Hilmar Lapp, the HP working group, Paul Morris, John Pickering, John Deck, the iDigBio staff, BRIT staff and others for their input on workshop, hackathon and alternative event inter-related content and direction as well as help with community outreach.
Social Values in a Future Internet: Analyzing the Named Data Networking Protocols

Katie Shilton
College of Information Studies, University of Maryland, College Park
kshilton@umd.edu

Abstract

The technical details of internet architecture affect social debates about privacy and autonomy, intellectual property, cybersecurity, and the basic performance and reliability of Internet services. Networking engineers making technical choices also weigh social values (consciously or un-) when working on internet infrastructure. This paper reports on ongoing research examining the social values considered in the design of a future Internet architecture, referred to as Named Data Networking (NDN). The project seeks to understand how a network of actors, work practices and contexts contribute to the embedding of particular values in NDN. This paper describes the values expressed by the engineering team based on data gathered from publications and internal documents. It uses these documents to understand the salience of particular values to the project. Values central to NDN design include efficiency and dynamism, which are invoked in response to technical constraints and possibilities. Other central values, including privacy and anonymity, stem from a concern for personal liberties. More peripheral communitarian values espoused by the engineers include democratization and trust. This taxonomy of values embedded in the current NDN approach can help advocates for values in design as well as policy researchers understand the social challenges and values engendered by this emerging architecture. The paper uses this taxonomy to discuss the implications of these values for information access, public policy, and existing institutional power dynamics.

Keywords: values in design, internet protocols, information policy, social and community informatics, qualitative data analysis

Acknowledgements: National Science Foundation, # CNS-1040868.
Copyright is held by the author.
Sociomateriality in the Sandbox:
Visually Oriented Documenting Practices of Young Entrepreneurs

Jamie Snyder
Syracuse University
jasnyd01@syr.edu

Carsten Østerlund
Syracuse University
costerlu@syr.edu

Abstract

Entrepreneurs are important drivers of innovation. Their work is highly creative while also requiring a distinct paper trail for processes such as incorporation of a business, filing of patents, and management of shifting resources. In order for entrepreneurial team members to interact in productive and creative ways, they need to establish shared and stable frames of reference, a sociomaterial field to support collaborative decision-making and coordination. This paper explores how young entrepreneurs build a sociomaterial field as they form their organization. In doing so, the authors draw on three bodies of work: document theory, presence/absence as articulated by John Law and visual materiality. Through exploratory analysis of an in-depth field study of a 12-week bootcamp for student entrepreneurs organized by a research university in the Northeast the authors describe three types of document oscillations emerging as the entrepreneurs build their sociomaterial fields: 1) Personal, Collaborative, Custodial; 2) Aspirational, Operationalized, Contractual, Standardized; 3) Temporary, In Progress, Stable. The approach developed in this paper contributes to the sociomaterial literature by highlighting the oscillating nature of visual materiality.

Keywords: documenting practices, visual information, sociomateriality, information behavior, social and community informatics, qualitative data analysis

Copyright is held by the author.
Highly Dedicated Enthusiasts: Online Museum Visitors Doing Everyday Life Information Seeking

Mette Skov
Royal School of Library and Information Science, Denmark
ms@iva.dk

Abstract

A current trend is to make museum collections widely accessible by digitising cultural heritage collections for the Internet. The present case study aims to add to previous studies by exploring what characterises online museum visitors in an everyday life information seeking context. Three factors are explored: information needs, information seeking and information use. A triangulation of research methods is applied. A web questionnaire survey gave initial, quantitative information about online museum visitors to a Military Museum. The follow-up interviews with 24 online museum visitors obtained rich, qualitative data to validate and elaborate the characteristic of online museum visitors’ information seeking behaviour. Based on the serious leisure perspective two different user groups are identified: collectors and liberal art enthusiasts. Across the two hobby classes, user study participants can be characterised as special interest museum visitors pursuing a long-standing interest or hobby. Participants’ dedication and often considerable time spent on the hobby indicate that their hobbies are integrated components of everyday practices. Although overlapping, the two hobby types have distinct profiles including the nature of knowledge acquisition. The identified information needs were surprisingly well-defined known item needs and only few exploratory information needs were identified. In relation to information sources and channels, participants stressed the importance of personal channels and the social context of the hobby. The present study contributes confirmatory to the research area of everyday life information seeking within serious leisure. The results also contribute to the emerging field of museum informatics by adding to the characteristic of the elusive online museum visitor, e.g., visitor motivation and the leisure context of the museum visit.

Keywords: everyday life information seeking; user study; serious leisure perspective; museums; cultural heritage

Copyright is held by the author.
Conceptualizing the Information Field: A Multidimensional Approach Based on Interdisciplinary Theory

Dorte Madsen
Copenhagen Business School
dma.ikl@cbs.dk

Abstract

The purpose of this paper is to explore the challenges of developing a shared conceptual framework for the Information Field, based on interdisciplinary theory. This paper argues that to study the Information Field, we need a multidimensional framework that includes disciplines and discipline-based assumptions. The paper builds on a continuum of integration from multidisciplinary to interdisciplinary to transdisciplinary, and a problem-based approach. The paper suggests a 3-dimensional matrix with material fields (information, technology, people) along a horizontal axis, analytic levels of disciplines along a vertical axis, and problems along a depth axis. The proposed framework lends itself to a nuanced analysis of the components and dimensions involved in multi-, inter- and transdisciplinary work. The 3-dimensional matrix offers a lens through which to view the contributing disciplines.

Keywords: interdisciplinarity, conceptual framework, information field, material fields, disciplinary components
Patient-Provider Communication and Information Access via Online Patient Portal: Understanding User Behaviors to Improve Healthcare and Outcomes

Xiaomu Zhou  

xmyzhou@rutgers.edu  

Si Sun  

sisun@eden.rutgers.edu  

Rutgers, The State University of New Jersey

Abstract

The patient portal is a relatively new healthcare information technology that enables patients more convenient access to their healthcare information and send messages to their physicians. Our study, based on 1172 communication threads (including patient-provider and provider-provider interactions), examined the themes discussed in these messages and different ways in which patients communicate with their providers via a portal employed in a large medical center. We also explored the differences between the patient portal and more traditional communication media (e.g. telephone, mail, and fax), and investigated the advantages and potential problems of the portal system. Further, we also analyzed how the new communication and information workload resulted from the adoption of patient portal was arranged among different clinical groups (nurses, secretaries, and physicians). Our findings show a wide variety of topics discussed in the communication messages (such as medication, appointments, laboratory tests, etc.) and how patients provide information, consult their providers, and express psychosocial and emotional needs. The work arrangement for additional information responsibilities among clinicians may also create new communication issues (e.g. information fragmentation, gap, and inconsistency). Finally, we addressed the implications on the organizational policy improvement, incentive structure, and a better design of the information system.

Keywords: patient-provider communication, health informatics, patient portal, EHR, human-computer interaction


Copyright is held by the authors.
Sociomateriality of Information Standards in Infrastructure: Structure, Organization, and Process Constraints

JoAnn Brooks  
Syracuse University  
jbrook03@syr.edu

David Mann  
The Mitre Corporation  
damann@mitre.org

Abstract

Information infrastructure comprises multiple disparate “worlds,” each of which may incorporate a distinct “regime of information” (Ekbia & Evans 2009). Different information regimes may be founded upon different types of information standards that do not interoperate with each other. In this paper, we develop a “sociomaterial” (Orlikowski & Scott 2008) perspective on information standards that illuminates such disjunctures through identifying fundamental distinctions between types of information standards: characterizing them in terms of structure, organization, and process constraints.

We start by surveying two groups of information standards: those designed for computational use, and those used in social ordering and organizational contexts. Because the separation between the two groups is unambiguous, we refer to these as “pure” versions of “machine oriented” and “human oriented” information standards respectively. We proceed to characterize fundamental differences between the logical structure(s) of these two groups in terms of differences between categorization and classification (Jacob 2004). Then we discuss key organizational processes integral to the management and use of each type of standard, and note fundamental constraints underlying differences between these organization processes.

We then move on to discuss “sociomaterial” standards – i.e. information standards the use of which entails imbricated combinations of computation and social practice. In addition to detailing structure and organizational process for each of these types, we argue that although these sociomaterial information standards are designed to bridge between formal computation and social practice contexts of use, fundamental distinctions remain such that all sociomaterial standards are ultimately either machine oriented or human oriented in their performance.

We briefly illustrate the power of our perspective with examples and provide a catalog of types of human oriented standard(s) which we believe to be the first of its kind. We conclude by touching on implications for sociomaterial theory development and noting practical implications for developing information infrastructure.

Keywords: information standards, standardization, sociomateriality

Acknowledgements: An earlier version of this paper was presented at the 7th International Conference on Standardization and Innovation in Information Technology (SIIT 2011). We thank Deb Bodeau, Richard Games, Martha Lorber, Leo Obrst, Anne Rawls, William Waters and anonymous reviewers from SIIT 2011 and iConference 2013 for helpful comments on earlier versions of this paper. Portions of this work were funded through grants from The MITRE Corporation.


Copyright is held by the authors.
Citizen Science Online Gaming: Work or Play?”
or
“Some of the Values in the Design of Web-based Citizen Science Games: The Case of Phylo”

Brian Beaton
University of Pittsburgh
bbeaton@pitt.edu

Abstract

This paper concerns citizen science. It investigates some of the values built into the design of web-based citizen science games, with a focus on a game called Phylo. Using methods from science communication studies and the interdisciplinary field of science and technology studies (STS), the paper offers a close reading and content analysis of a peer-reviewed scientific article written by Phylo’s creators. The article is entitled “Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment” (2012). In the article, the creators of Phylo explain how the game allows citizen scientists to assist with research in evolutionary biology under the guise of solving puzzles. Results of the content analysis suggest that Phylo’s creators have a negative view of wider gaming culture. The results also suggest that Phylo players are conceptualized as human resources. These findings will contribute to knowledge about citizen science game development. Understanding how the creators of citizen science games view gaming and gamers can help us understand the conceptual limits that exist for improving upon such games. This paper shows that some citizen science game developers view citizen science as a work formation.

Keywords: science, data, gaming, work, play education, history and philosophy of information, research methods, social and community informatics

Copyright is held by the author.
Abstract
Current search engines satisfactorily return relevant, ranked results to most posed queries. However, when searching on a dense topic for individual or collaborative learning purposes, the highest ranked results retrieved by these engines might not be the best starting point for learners given their current level of competence. We leverage concepts and computational solutions related to peer knowledge and interaction data in order to convert ranked search results in So.cl into sequenced results that allow learners to start with sources that are accessible and understandable before moving to increasingly advanced and complex content.

Keywords: social search, collaborative learning

Introduction
Searching for information on the Internet has become a daily activity for many of us, and over the years, search engine algorithms have greatly improved in their ability to retrieve applicable results to posed queries (Purcell, Brenner, & Rainie, 2012). Many of us do not even look further than the first few returned results to find a link with the information we desire (Jansen & Spink, 2006). However, if submitting search queries for learning purposes and not for the quickest, most succinct result, then current search engines leave some things to be desired. In such instances, the best and most relevant answer may not be desired immediately as the best answer may not be comprehensible to a novice. The preferred ordering of results, however, would first include results that contain the most basic, rudimentary understanding of the topic that match the competence level of the searcher before results that are more technical and complex.

Learning a topic with the help of a search engine may not be ideal for everyone. We do not deny that learning can be more effective and perhaps more enjoyable when working directly with others who can either organize the learning experience, immediately answer questions as they arise, or demonstrate the execution of tasks as in the case with over the shoulder learning, e.g., (Twidale, 2005), or with social search engines like Aardvark (Horowitz & Kamvar, 2010). Yet in the cases when learning in a traditional setting or even a non-traditional setting like a massive open online course or an after-school program is not an option, search engines combined with the social traces created in online social network communities like So.cl can help the learning process. Guided by the notions of Vygotsky (1978) and Lave and Wenger (1991) that social interactions are integral to learning, we introduce a way for the search engine embedded in So.cl to retrieve and present search results in a way that makes it easy for learners to start with sources that are accessible and understandable before moving to increasingly advanced and complex content.

The changes that we propose to the So.cl search engine will utilize data easily provided or calculated within the So.cl community including location, peer interests, and the past search behavior of both ourselves and our peers. These changes can be made using the affordances and information already available in the So.cl community and with minimal changes to the So.cl interface.
Related Work

Using social or collaboratively generated information to filter (e.g., (Shardanand & Maes, 1995; Bellogín, Wang, & Castells, 2011; Horowitz & Kamvar, 2010; Evans & Chi, 2008)) or retrieve information (e.g., (Kirsch, Gnasa, & Cremers, 2006; Twidale & Nichols, 2009; Hust, Klink, Junker, & Dengel, 2003; Morris & Horvitz, 2007; Pickens, Golovchinsky, Shah, Qvarfordt, & Back, 2008)) is not new. Twidale and Nichols (2009) and Hust (2003) discuss using information from users with similar interests and queries to aid subsequent users, which is a concept we propose would greatly improve the learning experience in So.cl when users perform web searches, especially if search queries include ambiguous phrases or location-specific information. Morris and Horvitz (2007) and Horowitz and Kamvar (2010) introduce SearchTogether and Aardvark respectively which both demonstrate how social and collaborative information can be utilized during search including the importance of the awareness and persistence of social data. And finally, Shardanand and Maes (1995) and Bellogin et al. (2011) discuss using or building profile data to infer the tastes of users to improve future recommendations. As a social network community, So.cl encourages users to enter personal information and preferences in their profiles and on their pages in the form of text and images. In the next section, we will discuss how we propose that this information can be used when retrieving and presenting users with search results to learning focused queries.

Learning Focused Search in So.cl

Individuals perform searches for different purposes including transactional, navigational, and informational reasons (Evans & Chi, 2008). This is probably not altogether different within So.cl. As So.cl is designed around learning about and through the interests of its community members, we would, however, anticipate even more searches that are informational. As such, supporting learning while allowing individuals to perform other types of searches as well would be opportune. So, as opposed to completely removing and/or altering the present searching functionality therein in hopes of improving the learning experience, we propose introducing two web search modes: Default and Learning (see Figure 1). As this searching is performed within the context of a social network community, the Learning search can take advantage of data that traditional search engines cannot. Below, we use an example usage scenario to explain the different types of data that can be used, how they can be used to order and assign a difficulty level to results (i.e., Beginner, Intermediate, and Expert), and ways that the results in the Learning mode will be presented that differ from the Default mode. The main factors that determine the ordering of results in Learning mode include location, the readability level of result pages, search term category, peer interests, peer link traversals, and explicitly assigned difficulty levels.

Example Usage Scenario

Jaime and Leslie are friends searching within So.cl to satisfy two different needs. Jaime’s car has mechanical issues, so she is searching for information on car engines while Leslie has learned that she is pregnant and is looking for information on baby showers and calligraphy so that she can learn how to make her own baby shower invitations. Leslie goes to So.cl and performs a series of searches on "calligraphy" (Figure 1) and "shower" while Jaime creates a post with links focused on "car engines" so that she can learn about car engines and share her findings with others (Figure 2).

Factors Utilized to Filter and Order Results in Learning Mode

Location: The location of So.cl members is explicitly provided in their profile information or can be ascertained through their IP addresses. This information can be used to return results that are geographically relevant. For example, garages in Jaime’s surrounding area or local trade schools offering courses on car repair would be returned as she looks for information on car engines.
Figure 1. Results returned in both the Default and Learning modes when searching for "calligraphy".

(a) Default mode
(b) Learning mode

Figure 2. A user-created collection of links in which the user has the ability to manually assign difficulty levels to each link.

(a) The current screen presented when making a post on "car engines"
(b) The modified screen when making a post on "car engines"

Readability level: The language used in webpages returned by searches can be a hindrance to learning if the readability levels, including vocabulary and sentence structure, are high or if they include jargon that one may not understand as a beginner. An example of this is the term "letterform", which Leslie encountered while searching for information on calligraphy. To aid with the learning process, such webpages would be automatically assigned higher difficulty levels than others with more straightforward, less jargon-filled writing.
**Search term category:** Not only should the readability level of the webpages returned influence order, but the type of search term should as well. Technical topics, for example, should yield results that look different than cooking topics. Results assigned a beginner difficulty level in cooking searches would probably include lots of images and videos and includes lists or steps while beginner technical results would probably include code snippets and more explanatory text.

**Peer interests to help with term disambiguation:** There are some instances where search terms may have the same spellings as other words. For example, “tree” can refer to the plant or the data structure. “Python” can refer to the programming language or the animal, and likewise, when Leslie searches for “shower” instead of the more specific phrase “baby shower” So.cl may not know what types of results to return. Fortunately, Leslie’s friends have also recently had children and have been performing searches and building collages within So.cl looking for interesting baby shower gifts, so So.cl uses that information to help with term disambiguation and returns information about baby showers and not bathroom showers.

**Peer link traversals:** Awareness of what others have done and have seen can also help during collaborative web searches (Morris & Horvitz, 2007). Links that have been visited multiple times by peers in So.cl and the order in which they were traversed, especially if that path was repeated multiple times, can be utilized to help order result links for others.

**Explicitly assigned difficulty levels:** Morris and Horvitz (2007) note that participants in their study requested the ability to be able to manually edit and order pages visited into an order that had meaning for them. We propose allowing So.cl members to explicitly assign difficulty levels to search results included in posts (see Jaime’s post in Figure 2). These explicitly assigned difficulty levels can then be used in addition to all of the other information when assigning that particular resource with a difficulty level and to organize and filter results for subsequent searches that include that resource.

**Changes to the So.cl Interface**

All of these changes can be implemented with minimal changes to the So.cl interface as seen in Figures 1b and 2b. The results when searching in both the Default and Learning modes can be displayed in the same place and toggled with a link (Figure 1). The biggest change seen in Figure 1b are the initials B, I, and E along the left of the screen. These designations refer to the difficulty levels of the links: Beginner (B), Intermediate (I), and Expert (E). Links in between the initials have the same difficulty level of the link preceding it (e.g., the second link in Figure 1b is a beginner link). Individuals can choose to view only the links of a particular difficulty level by choosing that level in the dropdown menu displayed when clicking on the downturned arrow to the right of the Mode options. Learners are expected to begin with the links at their current, self-identified competence level, yet they are not prevented from starting at another point if desired.

Allowing individuals to explicitly assign difficulty levels to resources can also be performed with small changes by having the “Comment|Tag|Add Link” functionality applied to each resource in a post (Figure 2b) as opposed to the entire post (Figure 2a). The ability to assign the difficulty level to each item could then be added to that list of options with a similarly styled “Suggest difficulty level” link.

**Conclusions**

We present a way to encourage and support learning focused searching within So.cl that involves presenting search results in a way that allows learners to start at their present competence level before moving on to more advanced content. The factors that we reference – location, peer interests, result readability level, type of search term, peer link traversals, and explicitly assigned difficulty levels – can all be utilized to filter and determine which results to return to learners and the order in which to present them. Given the interest driven focus of So.cl and the type of data collected therein, this can be easily implemented within the social network community without major changes to the experience or to the interface.
References


K-Pop Live: Social Networking & Language Learning Platform

Chin Leng Ong
cli.ong2010@sis.smu.edu.sg

Kian Ming Png
km.png.2010@sis.smu.edu.sg

Aloysius Lau
pohchyelau.2010@sis.smu.edu.sg

Houston Toh
houston.toh.2010@sis.smu.edu.sg

Feida Zhu
fdzhu@smu.edu.sg

K-Pop live is a social networking and language learning platform developed by an undergraduate student team from Singapore Management University. K-Pop live aims to combine social media together with gamification to promote Korean culture. It consolidates all relevant Tweets from Twitter as well as videos from YouTube. The platform allows the user to connect with his friends who share similar interests in terms of K-pop artists and music.

**Keywords**: social network, K-pop, language learning, gamification

**Abstract**

K-Pop live is a social networking and language learning platform developed by an undergraduate student team from Singapore Management University. K-Pop live aims to combine social media together with gamification to promote Korean culture. It consolidates all relevant Tweets from Twitter as well as videos from YouTube. The platform allows the user to connect with his friends who share similar interests in terms of K-pop artists and music.

**Keywords**: social network, K-pop, language learning, gamification

**Introduction**

K-pop live is a branch of PalanteerRT (http://research.larc.smu.edu.sg/palanteer) developed by Living Analytics Research Centre (LARC) at Singapore Management University. It is a live real-time application based on a state-of-the-art data mining technology that collects relevant and trending data from major social media sources such as Twitter and YouTube. Our objective is to develop a system to transform LARC’s research into a feasible consumer application. K-Pop live is an interactive iPad application where users can easily tap into their favorite K-Pop artists and connect with friends who share similar interests. The platform updates the users by continuous analysis of their surfing behaviors and interests. Through the platform, users can acquire the latest K-Pop news and media through a consolidated newsfeed. It also allows users to share contents with their friends through integrated Twitter and YouTube features. The platform is equipped with gamification features designed for language learning the fun way. It recommends K-Pop songs to users based on their interests and prompts them to participate in “Learn Language” game. Each song is rated differently based on the level of difficulty (community rating). Successfully completing a song is similar to completing a quest in games. Users acquire points and get recognized for their achievements.

**K-Pop Live’s Features**

In the following section, we describe features of K-Pop live. The last decade has seen an explosive growth of online social networks. Of the widely popular social networking platforms is Twitter. If you are like Houston, who has interests in many things, soon you will find it difficult to sort out and organize tweets by your specific interest topics or categories. Houston, an undergraduate student, is seeking to network with those that share similar interests in music. K-Pop (Korean popular music) has gained significance fan following all over the world along with K-Drama (Korean drama and movies). Houston is...
planning on doing a semester of exchange in Korea and he is seeking a fun, interactive way to learn more about Korean culture and language.

Profile Management

In My Profile page, users can edit their profile details. Users can scroll through all K-pop artists and select K-pop artists of their interests. The platform will continuously monitor online popular K-pop artists and make this list updated. The platform will record these selections and in the background, it will sort out and organize tweets and YouTube videos based on these selections. Users can go back and add or remove K-pop artists of their choice from their “Favorites” list.

Custom Newsfeed

In My Profile page, users can edit their profile details. Users can scroll through all K-pop artists and select K-pop artists of their interests. The platform will continuously monitor online popular K-pop artists and make this list updated. The platform will record these selections and in the background, it will sort out and organize tweets and YouTube videos based on these selections. Users can go back and add or remove K-pop artists of their choice from their “Favorites” list.

Profile Management

K-Pop live displays up to six news items (tweets or YouTube videos) in each page. Users can flip through the pages on the screen. Click on each news item will expand that particular item and show the entire content. Users can re-tweet or tweet right from the screen. If it is a video, users can directly play the video from the screen. If it is a website, users can directly view the website without leaving the K-Pop live platform.

Forming Interest Groups

Of the many users you follow and many users that follow you in Twitter, how do you know to tune into whom for which kinds of news? Likewise, of the many users in YouTube whose videos or channels you follow, how do you know to tune into whom? K-Pop live makes it easy to form interest groups. This way, you can further refine tweets and YouTube videos based on shared interests. Additionally, this allows you to easy share your tweets or YouTube videos with only those that are interested in the K-Pop artists or music features in the media. Users can easily create, remove, or edit interest groups in K-Pop live.
Language Learning Game – Overview

Viewing, re-tweeting, and sharing your favorite K-pop news and videos is cool. K-Pop live monitors users’ browsing behaviors and uses it to recommend K-Pop songs for language learning. Each song is rated by the community based on its difficulty. We selected top K-Pop songs viewed and played from K-Pop live and a panel of Korean speakers reviewed and assigned initial difficulty rating. This panel of Korean speakers also contributed translations as well as English phonetics, which will be used in our game. We plan on bringing more crowd sourcing into this process going forward, while maintaining a panel of Korean speakers from the community to mediate.

Language Learning Game – Game Mechanics

Users start out as Level 0 (not having played this game). They will be prompted to participate in language learning game. They can choose to ignore this and turn off this feature permanently, if they choose to. Based on the K-Pop artists from their Favorites list and their browsing behaviors, they will be given a list of songs to choose from. If this is the first time a user is playing this game, all songs will be at Level 1. Users can choose a song and “unlock” it by successfully singing along.

The platform will measure two aspects of singing. First, it will perform pitch detection. Second, it will perform pronunciation detection. Each song is divided into multiple verses. Scoring will be done for each verse and the composite score is the average across all the verses. For users to successfully pass the song, they must score beyond a set ratio (for now, it is set to 50%). Upon successful completion, users will acquire XP (experience points). All songs at a particular level are assigned a set XP. Supposing that Gangnam Style (by Psy) is assigned 100 XP, Houston scores 100% on pitch and 50% on pronunciation. The resulting mark will be (50XP * 1.0) + (50XP * 0.5) = 75 XP.

Once users reach a certain XP mark, they will level up. As of now, with K-Pop live platform, each level is assigned a fixed XP. From Level 1 to Level 2, it takes 500 XP. From Level 2 to Level 3, it takes 1500 XP and so on. It is designed in such a way that it is increasingly more demanding to level up. However, more “difficult” songs come with higher XP.

Language Learning Game – Social Sharing & Social Playing

Users can play and practice each song as many times as they want. When they are ready, they can click on “Challenge” and their singing will be recorded. Pitch and pronunciation detection will then be done and the final marks will be released on the screen. Users can then publish this result to Facebook, tweet it in Twitter and share with the world.
K-Pop live connects users to other users through singing battles. It matches up two players based on shared interests in K-Pop artists and music. As of now, we match up two people if they share similar K-Pop artists and music as well as same game level. There are two ways to participate in this singing battle.

Users can choose to participate by clicking on “Challenge Me!”. K-Pop live will then go and look for active users online and prompt them to participate. Or, users can turn on “Other users may challenge me” option and be online. Then, K-Pop live will connect the users to others opting for “Challenge Me!” option. Users must accept the challenge within three minutes, otherwise the offer expires. Participating in singing battles will lead to gaining of bonus and “XP boost”.

Upon two users accepting the challenge, singing battle will begin. K-Pop live will perform pitch and pronunciation detection in much the same way it would for individual singing challenge. It will then compare the score and display the winner/loser on the screen. The winning player will gain both bonus and XP boost. The losing player will gain bonus but not XP boost. They can choose to publish the singing battle result to their favorite social networking platforms. They also acquire “Social Badge” by participating in singing battles.

Whenever there is a new song released in the game, K-Pop live will add “XP boost” option to it and advertise. We envision running different kinds of campaigns to get more users to participate and get their friends to come and participate (that may not necessarily be already playing our language learning game).

Language Learning Game – Language and Culture Learning

K-Pop live features Korean Alphabet & Vocabulary feature. Users can click on words in song lyrics and K-Pop live will display a pop-up window showing the chosen word in Korean (Figure 3). It shows the breakdown of the word in terms of consonants and vowels. It shows how the pronunciation is determined.

Singing along and challenging friends can be fun. Upon completing each song, users are prompted to participate in a Quiz. It is a phrase matching game where users will be prompted to match Korean-written verses to English Phonetics to English translations. Five verses from each song will be given and each correct mapping will result in a bonus.

Additionally, music is a powerful way to connect cultures. Clicking on phrases or words in song lyrics will pop up Information window, which will show news or Wikipedia entries related to the selected phrases or words. Right from K-Pop live platform, users can look up all information related to the song, song artist, or any keywords or news relevant to the song.

Figure 3
Conclusion

K-Pop live aims to be a one stop portal for all things related to K-Pop. With its various social networking features geared towards customized newsfeed and news sharing, users can now easily access real-time news about K-Pop catered towards their specific interests. Using this user preference and browsing information as well as social information (e.g. who shares similar interests with who), K-Pop live makes language learning fun and relevant. Through gamification, users can learn Korean language the fun way and learn it together with friends who share similar interests as them.
Cairn: Using Digital Markers to Bridge Learning Communities

Amanda Menking
amenking@uw.edu

Ivette Bayo
ibayo@uw.edu

Sean Fullerton
sfullert@uw.edu

University of Washington

Abstract

Cairn is a proposed geo-social networking mobile application that could help people search for, find, participate in, and tag learning communities, events, and artifacts. The goal of Cairn would be to enable users to find others with whom they can study, attend learning events, and share learning related artifacts, such as museum exhibits or historical markers. The primary target audience would consist of post-secondary students; the secondary target audience would be recent graduates and young professionals. Cairn is proposed to be a native mobile app integrated with existing web-based interest and/or social networks. This paper presents a design exploration, including an overview of the relevant sections of the Information and Communications Technology (ICT) geo-social and learning markets, eight key features identified by focus groups and interviewees, a case informed by user research, and possible directions for prototype development and integration with So.cl.

Keywords: mobile app, interest network, geo-social network, informal learning, post-secondary education

Introduction

Cairn is a proposed geo-social networking mobile app that would help people search for, find, participate in, and tag learning communities, events, and artifacts. Cairn would enable users to find others with whom they can study, attend learning events, and share learning related artifacts, such as museum exhibits or historical markers. Cairn is proposed to be a native mobile app, not a web-based app; however, it could be integrated with an existing web-based application like Microsoft Research’s So.cl.

The primary audience for Cairn would be the post-secondary student. Students attending large community colleges and universities often have difficulties finding appropriate resources to improve their learning experiences outside the classroom. Undergraduates can feel lost in a sea of students (especially in large introductory courses), overwhelmed by available opportunities, and unsure about how to make connections that will facilitate their academic success.

Cairn’s secondary audience would be the recent graduate or young professional. Recent graduates entering the work world may find their social networks significantly limited compared to the contacts they enjoyed during their student days. They may also be transferring knowledge from an academic to a practitioner ecosystem and need support in doing this (Candy & Crebert, 1991).

ICT Geo-Social and Learning Market: Brief Overview

Geo-Social Networking Apps

The Information and Communications Technology (ICT) geo-social market currently consists of popular check-in enabled apps such as Foursquare and Yelp, lesser-known apps such as Loopt Mix, WhosHere, Friendthem, and Groupie, and dating apps provided by online sites like OkCupid, Blendr, and Grindr.


Copyright is held by the authors.
Additionally, Meetup offers both a website and an app designed to enable users to connect with one another based on shared interests. Facebook offers Places and integrates with several existing mobile geo-social networking apps such as Instagram.

None of the services listed above provides a user the capability to find another person on the basis of shared academic goals or specific learning interests, nor do any of these services enable a user to find a specific learning event or artifact. Meetup comes closest, but it is not designed to interface directly with a user’s registration in a formal learning institution, such as a college, university, or professional program.

Educational Apps

Currently available educational apps may be divided into games (e.g., electronic flash cards and puzzles), mnemonic aids, and subject-specific guides. That is, most educational apps focus on content rather than learning experiences or opportunities. Moreover, the majority of these apps are geared toward younger users; one study found that over 80% of educational apps available in the iTunes App Store are created for users between the ages of toddlers and high school students with all educational apps for young children exceeding apps for adults by almost 20% (Schuler, 2012). Educational apps geared towards older teens and young adults remain elusive.

An Empty Space: Geo-social Networking and (Informal) Education

Cairn would fill a gap in the geo-social networking and educational apps markets. Although many apps using geo-location services enable people to meet one another and even to create communities around shared interests, very few explicitly tie formal learning objectives to informal learning opportunities. The exceptions include:

- In 2011, Arizona State University deployed an app similar to Cairn specifically for their online students. Spark enables online students to arrange face-to-face meetings with other ASU Online students and staff to discuss homework and assessment in a digital space accessible only by members of their immediate learning community (Kawam, 2011).
- In February 2011, OpenStudy was released from beta to production, quickly registering tens of thousands of (mostly secondary) students. Using principles of gamification, OpenStudy enables students all over the world to create profiles, ask questions, study subjects, and earn badges (Tsotsis, 2011). OpenStudy has recently begun to offer enrollment in massive online open courses (MOOCs) and grant related certifications for a fee.
- Also in 2011, Ryerson University Computer Science students developed Longitude, a web-based and mobile app that enables students to find each other based on proximity, communicate with one another via chat, phone, and video, and create shared, tagged maps (“Student-developed geo-social,” 2011).
- In 2012, students at the University of Washington working with the university's IT department developed SpaceScout, which helps UW students find study spaces based on selected criteria.

These apps, however, are either limited to specific institutions’ online learning communities (Spark), focus exclusively on online relationships (OpenStudy), have not yet been released (Longitude), or meet more finite requirements (SpaceScout) than Cairn would. What would set Cairn apart is its emphasis on tying post-secondary academic goals and informal learning opportunities to a user’s interest networks in real life.

Assessing and Meeting Market Needs

User Research

We held three focus groups and nine one-on-one interviews with undergraduates and graduates at the University of Washington and undergraduates at North Seattle Community College to discuss the
prospects for an app such as Cairn. Students expressed interest but raised issues about privacy, information overload, and integration with existing applications. Further probing about these concerns provided guidance for a next stage in exploring design considerations and developing a use case.

Design Explorations

From our interviews and focus groups, we identified eight design features that students believe would create utility for Cairn users. These are summarized below in order of importance.

Authentication

Using an existing authentication system would increase a user’s sense of privacy and protection. For example, in the case of the University of Washington, only users with a UW Net ID would be permitted to connect with Cairn services related to UW courses and students. Ideally, Cairn would work with a variety of authentication systems including those used by tertiary learning institutions and professional networks.

People Finder

Essential to the success of Cairn would be the capability to find other users within one’s proximity. By finding people who are interested in and studying the same subjects, a user can broaden his interest and social network, help to contribute to and invest in a learning community, and move one step closer to meeting his personal and professional learning goals.

Event Finder

The capability to find events, such as speaker series, museum talks, plays, and performances related to one’s learning objectives would also be essential to the success of Cairn. However, as noted by several of our interviewees, this feature is most important when ample notice is given and a calendar integration feature is available.

Calendar Integration

Many of our interviewees were interested in calendar integration. Several said this would be one of the most valuable features as they currently have challenges managing all of the events they hear about via email, listserv, and word of mouth.

Opt-in and Opt-out

Being able to opt in and out of Cairn at different times would be an important feature because students want to control when their activities may be viewed by classmates, other students, or other professionals. Opting out not only protects one’s privacy but also provides for the control of information flow. The opt-out feature would be universal in that all geo-location notifications would be turned off during opt-out times.

Artifact Finder

Introverted users, in particular, have indicated that they might like finding and tagging artifacts, such as museum exhibits, historical markers, botanical gardens, and artwork. The learning opportunities offered by artifacts can easily be overlooked; however, some students and instructors have indicated interest in connecting course content and discussions with related items beyond the classroom. Additionally, one interviewee, who currently works as a wine steward, said she would find this feature extremely helpful in her professional life and interest network.

Tagging and Commenting

A feature that enables users to comment upon and tag events and artifacts would enable personalization and increase engagement. Collaboration and contribution would be important to the robustness of an app like Cairn.

Push Notifications

When enabled, push notifications could serve two purposes: to alert users of people, events, and artifacts within proximity, and to remind users of the benefits of Cairn even when it is not open.
Use Case

Nico is a University of Washington freshman currently enrolled in Economics 101, a large lecture class with over 200 students. Nico’s professor mentions an app called Cairn that enables her to connect with other UW students studying economics. Nico downloads the app and opts-in. The next time she is studying at Starbucks on the Ave, she receives a notification that three other students in the cafe are also studying economics, and they are willing to meet others to study together. Nico uses Cairn to message all three students. Two students reply, identifying themselves, and ask her to join them. They happen to be studying economics together and are happy for Nico to contribute.

After her study session at Starbucks, Nico walks back to campus. As she passes the Henry Art Gallery, she receives another notification from Cairn. The app tells her that a current exhibit includes a display discussing the economics of fine art print sales. Nico decides to look at the exhibit and tags it as being of interest to other students. Before she leaves the gallery, she receives a notification about an upcoming lecture, and she adds this to her calendar. She then opts out of the app for the next 12 hours so that she will not be disrupted during her afternoon classes.

Conclusion

Most young adults, especially females, feel more comfortable with mobile technology and social media than they do with traditional ICT solutions (Lenhart et al., 2010). Because mobile phones are ubiquitous, Cairn would afford the opportunity for the targeted audience to use the technology with which they are most comfortable to broaden their learning and interest networks. Cairn would also provide an opportunity for users to connect formal learning objectives with informal learning events and experiences. Cairn would not provide educational content but rather draw attention to and enable young adults to connect with learning opportunities that already exist in their communities, but which might otherwise go unnoticed.

So.cl provides users an opportunity to create connections and communities based on shared interests rather than existing relational networks. Cairn, too, would build on the idea of interest communities by extending learning beyond the classroom and a set of structured objectives to informal opportunities and personal interests. A possible integration with So.cl would allow Cairn users to leverage their existing interest network on So.cl, employ Bing to find additional learning opportunities, and share their findings with an extended interest network. This kind of integration may also increase user participation and engagement for both So.cl and Cairn.

References

Abstract

With exponential increases of information flows on social media platforms, user experiences have become fragmented, cacophonous, and often overwhelming. Such "noise" makes learning on these platforms difficult. By contrast, the principle of polyphony frames learning with social media as a graceful, intentional threading together of multiple voices. In this paper we describe AmpDamp, a design concept for orchestrating polyphony in social media environments. Connecting a physical knob to a browser extension, AmpDamp translates the gestures of turning up (amplifying) and turning down (dampening) into a dynamic orchestration of the “volume” of a user’s social media feed. By focusing on issues of granularity and temporality, we position polyphony as an entry point into incorporating the values of control and curation into learning experiences using social media.

Keywords: polyphony, design, values

Introduction

With exponential increases of information flows on social media platforms, opportunities and sites for peer-based, informal learning abound. However, as information streams become floods, information competes for our limited attention and user experiences have become fragmented and often overwhelming. In this chaotic information environment, leveraging social media to establish and support informal learning proves increasingly difficult.

The issue here, and the motivation for our design exploration, is that current binary options of control over social media render users’ preferences as static configurations. Web services such as Facebook and Google enable hiding or viewing streams of content, while other curation tools use algorithms to condense content based on user-designated interests. Microblogging platforms such as Twitter barrage users with an unmediated stream of temporally organized content, ignoring volume or relations between a user and her network. By offloading choice to algorithms, these mechanisms force users into preferences that lack social nuance. We believe that such tools are inadequately flexible to support the wide range of situations and communities where informal learning occurs.

In order to better harness the interactivity and engagement afforded by social media platforms, we propose AmpDamp, a conceptual interface that gives users a range of filtering options for modulating Twitter streams in real time and across contexts. By doing so, AmpDamp supports users’ efforts to more effectively learn from their social networks. In creating AmpDamp we have foregrounded the concept of polyphony as a design principle. The metaphor of polyphony, taken from the arts, offers a lens through which we address questions of control and curation in informal, social learning environments.
From Polyphonic Works to Polyphonic Design

Polyphonic works are compositions that incorporate multiple voices without the presence of an authoritative voice. In polyphonic music or literature, instrumental or textual voices are interwoven throughout a composition while remaining rhythmically and melodically autonomous. Bakhtin (1984) argues that in literature, “the essence of polyphony lies precisely in the fact that the voices remain independent and, as such, are combined in a unity of a higher order than in homophony […] the artistic will of polyphony is a will to combine many wills, a will to the event.” In foregrounding polyphony as a value in design, we follow Bakhtin in his metaphorical use of polyphony. Bakhtin argues that polyphony offers a powerful metaphor to draw our attention to new challenges—from navigation to interpretation—that emerge when narrative expands beyond a normative monologic unity or when a musical composition features multiple simultaneous voices. Social media platforms present a parallel challenge, as they create noisy information environments for users; existing tools for managing content are inadequate to deal with the presence of a wide range of voices, rhythms, and volumes in a nuanced way.

As a framing metaphor, polyphony recasts the information overload problem as a question of how a diversity of voices may co-exist in real time. We offer an alternative to the formal, algorithmic, and static filtering of information streams by enrolling the user as an active participant in the dynamic modulation of her own experience. In doing so we locate our work within a community of interdisciplinary scholars focused on accounting for the role of human values in the design and use of socio-technical tools or systems; work in this space includes reflective design (Sengers et al. 2005) or values in design (Nissenbaum 1998, Flanagan et al. 2005). Sengers and Gaver (2006) argue for a nuanced, reflective understanding of design and use, one which eschews a single preferred interpretation of the functions of a given interactive system in order to acknowledge the co-existence and legitimacy of multiple - even competing - interpretations. What we’re calling polyphonic design argues for the need to design with polyphony as a guiding value and further to design flexibly, in support of multiple unspecified outcomes. Polyphonic design moves away from the idea of optimizing on a value such as polyphony or tightly engineering a polyphonic experience; rather than designing to produce polyphony, our work aims to enable a user to experience it.

Learning with Social Media

Learning with social media can happen in many different ways. Many researchers agree that informal, or learner-directed, actions on social media platforms can enhance and support in-classroom instruction (Dunlap & Lowenthal 2009, Richardson & Swan 2003). Most research that aims to facilitate learning with social media focuses on engaging learners to participate in an online conversation. However, moving outside of classroom settings there is little research that engages with how independently motivated users navigate social media content streams and learn from them in context.

The primary way a user encounters new social media content is by reading news feeds; these content streams are at times crowded with many competing voices or at other times quiet, based on the current happenings in a user’s network. Networks, in turn, vary across contexts and relationships, from friends and family to colleagues and professional acquaintances to entertainment and news. Each community in which a user actively participates adds social data. Parsing meaning from these diverse information sources requires a considerable amount of attention in order to contextualize content.

In both structured and informal settings, the value of attention—separating signal from noise, focusing on specific elements of the environment—grows as information is produced and circulated in new and unpredictable ways online. At the same time, a user’s ability to pay attention is strained by the scale, speed, or medium of information engagement. In social media contexts, where both time and attention are limited, tools for filtering the online experience are key to both targeted information search and serendipitous learning opportunities. Dunlap and Lowenthal (2009) argue that social media tools such as Twitter facilitate informal and just-in-time interactions that support and often extend traditional classroom-based learning. However, the control mechanisms we have for managing these timely social interactions are limited and do not sufficiently support varying levels of engagement.
Listening to Social Media

Crawford (2009) has written extensively on “listening” as a metaphor for considering “the nuances of connection and communication” afforded by online social media platforms. Crawford describes three modes of listening online: reciprocal, background, and delegated. Reciprocal listening implies a conversational approach, such as responding to @mentions or requests. Delegated listening outsources bits of attention to automated listening “assistants” such as news alerts. Finally, Crawford compares background listening to the ambient noise of the radio, with the lowest level of attention possible. Crawford argues that it is important to pay attention to the labor inherent in listening. Important to Crawford’s argument is that boundaries between these different kinds of listening - and the degrees of attention they require - are fluid, and users are likely to switch between modes throughout the day.

Polyphonous Design: Learning with AmpDamp

AmpDamp is a proof-of-concept for enabling the experience of polyphony in social media environments. Current binary options in social media interfaces (follow/unfollow, mute/unmute) fail to provide sufficient granularity for a user to manage the nuanced voices and rhythms in her social media streams. By combining hardware (a knob) with software (a browser extension), we propose a design to increase the granularity and temporal modularity of social media interfaces.

For example, Jan is a recent college graduate who has just gotten her first job as a programmer for a large technology company. Her new position requires knowledge of Javascript, a programming language she is just beginning to learn. After attending a Javascript meetup, Jan adds several new acquaintances—all of whom already know each other—to her Twitter feed. One of them, John, tweets a lot of interesting and useful links about Javascript. Unfortunately, he also tweets a lot of images of his cat, Felix, and engages in cat-related conversations with his programmer friends that appear as extended inside jokes to an outsider such as Jan. While Felix is quite charming, his near ubiquitous presence on Jan’s Twitter feed challenges both her motivation and ability to use John’s knowledge as a Javascript learning tool. The situation is further complicated by the fact that John tends to tweet Javascript-related material during traditional work hours, spending his non-working hours chronicling Felix’s movements; by contrast, Jan focuses her Javascript learning efforts on the evenings and weekends. Jan is left with few options: unfollow John and his friends and miss an opportunity to learn more about Javascript, or continue following John and try to ignore the copious cat pictures. The problem here is filter failure; Twitter’s current filtering features are too coarse to accommodate the dynamic nature of human interaction (Shirky 2008). On one hand, Jan learns a lot from reading the links about Javascript programming in her feed. On the other, sometimes useful information gets lost in a sea of cat-stagrams. What Jan needs is a tool to manage the voices in her stream that allows for varying levels of control and that recognizes that her preferences change over time. This example highlights two primary design concerns for supporting informal learning through social media interfaces, addressed in depth below.

First, the granularity of control in social media interfaces is too coarse to harness the affordances of social media in support of peer-based learning. Classroom-based learning tends to be structured, controlled, and formal; attention is focused on a single speaker, or on the linear progression of an ordered conversation among several voices. By contrast, learning in social media environments tends to be unstructured, opportunistic, and informal, with a multitude of voices present at any given time. As illustrated in the example above, binary control choices within existing platforms (on/off, follow/unfollow, mute/unmute) occlude a vast horizon of possible interactions and mute the dynamism of social media environments. As a proof-of-concept, AmpDamp consists of a variable resistor sending input to a custom Chrome extension through an Arduino microcontroller and locally running webserver. When a user visits Twitter.com our extension maps the knob’s signal to the CSS opacity and zoom attributes of DOM elements matching selected Twitter usernames or hashtags. This first instantiation functions as a very rough proof-of-concept operating only on Twitter, with custom built knobs and hard-coded selections. Our next iteration uses commercial knobs, eliminates the local web server (using Chrome’s USB API), and permits dynamic user selection of features for manipulation (e.g. usernames, hashtags, keywords).

Second, social media users’ listening needs and preferences change dynamically. When listening to the radio, a user modulates volume depending on the content of a broadcast or its relevance to what else she happens to be doing. In order to address the temporal dynamics of user needs and control preferences with respect to social media, we introduce a similar mode of interaction through the
combination of a physical knob and a contextually aware browser extension. This analog volume control knob provides a "natural" mapping between a physical action (twisting and turning dials) and an effect within a social media context, enabling users to quickly and easily modulate their Twitter streams using a deeply familiar interactional modality. The knob communicates user actions to the browser extension via USB, and the extension combines that input with contextual information based upon the currently active social media platform. This design enables the user to modulate social media preferences on the fly without the cumbersome work of configuring rules, profiles, or triggers. The design emphasizes the active and ongoing modulation of a user's social media environment over pre-configured sets of rules.

While dampening cat pictures might seem like a trivial example, it is meant to highlight the fact that AmpDamp's functionality is 1) user configurable and 2) context dependent. Furthermore, behavior may be fully customizable through a plug-in system. As designers, we do not want to dictate the repercussions of “amplify” or “dampen.” Rather, we want to enable users to construct those meanings themselves, and to adjust them over time and situations. Because we are still in the early stages of development and implementation, we continue to consider questions about how filtering criteria – and specifically the complex set of joint conditional probabilities that may arise in use – can be addressed technically. To this end, we plan to supplement our preliminary, paper-based user testing with further prototyping and user research; topics will include exploring how users navigate and cognitively manage multiple information needs and filtering criteria, and identifying specific learning heuristics and levels of control granularity that users desire or find lead to meaningful learning experiences.

Conclusion

We have created AmpDamp through an iterative process of design exploration, prototyping, and user research. By modulating social media voices with a hardware knob, AmpDamp addresses the inherent challenges of designing for, and with, polyphony within the scope of social media. Polyphony provides a rich entryway for designers to rethink the experience of social media platforms such as Twitter; further, it creates a space for physical control tools that may better reflect values such as granularity, temporality, and dynamic situational specificity.

We believe that by giving a different kind of control to users, AmpDamp can foster self-directed learning as well as promote a conscientiousness toward curation that current social media platform controls do not afford. By drawing attention to the often invisible work of filters and algorithms, we reposition filtering in terms of ongoing practice. Instead of trying to optimize around a concept (such as polyphony) or offload preferences to preconfigured technological artifacts, AmpDamp allows users to improvise and learn from their social networks in novel ways and in real time. Moreover, this reframing causes meaningful reflection on what values within social media might best support informal, peer-based learning. This final point—that we seek to foster active reflection on values at the point of design as well as in use—provides a direction for thinking about social media as a platform to support ongoing learning.

References


Towards a Methodology of Virtually Augmenting a Knowledge Sharing Community of Practice: A Case Study of the Local Food System of Denton, Texas

Andrew J. Miller
mail@andrewjm.co
Laura A. Pasquini
laura.pasquini@unt.edu
Leila A. Mills
leilamills01@gmail.com
Mark Evans
markevans@my.unt.edu

University of North Texas

Abstract

Observation, interviews, and a pilot study were conducted to identify requirements for social media to support knowledge sharing that will foster organization and growth of the Local Food System (LFS) in Denton, Texas. Due to lack of infrastructure, participants of this grassroots, self-organized, self-sustaining community of practice currently do not have access to physical or virtual space needed to share, retrieve, and archive LFS knowledge. Proposed system design and implementation will be based on completion of a mixed methods community study. Ethnographic techniques for a close-to-home investigation will examine how the existing LFS community operates. The goal is to design an unobtrusive LFS knowledge sharing social media site that is optimized for thoughtful archiving, accessible information retrieval, and lifelong community learning.

Keywords: media, knowledge sharing, community of practice, local food system, ethnography

Executive Summary

Introduction

The Local Food System (LFS) of Denton, Texas is composed of passionate individuals who make up a self-sustaining community of practice. It is a community with no clearly defined leader, which has grown organically out of pre-established agriculture practices in North Texas and with the rising national interest in organically grown, locally sourced food. Individuals and organizations bring varying specialized skills to the community that sustain the LFS. They communicate through a variety of channels, sharing information and knowledge integral to the community. There exist several organizations, such as the Denton Community Market, that strive to integrate the LFS community with the encompassing Denton community in order to bring awareness to the LFS. These organizations tend to communicate knowledge that encourages participation. Currently, no one is collecting all the knowledge the LFS relies on into one accessible interface. Knowledge within the community tends to be organized modularly with no infrastructure to facilitate an accessible outlet for shared community knowledge. It is not uncommon to be a member of the LFS community and not know about the existence of organizations that contribute integral knowledge to the community. In short, the knowledge base for the LFS community is disorganized, which is reflected in community awareness and growth.

Members of the LFS facilitate strong community bonds. They address issues of environmental sustainability by emphasizing native plant life and shorter delivery distance of food, decreasing fossil fuels burned during delivery. They also stimulate local economies through the production, exchange and consumption of their goods. Currently, the LFS community awareness and growth is limited by the lack of

Acknowledgement: Jeremy Voss and Kati Trice for your time discussing in detail your experience in the community.
Copyright is held by the authors.
accessibility to community knowledge. An individual must invest substantial time and energy into becoming a part of the community in order to learn the ins and outs. As a result, members tend to be very active in the community and share passionate interests in local food, while much of the population of Denton seems to remain largely unaware or unwilling to participate.

Organizational Background

Denton’s LFS is a grassroots, self-organized, self-sustaining, community of practice woven together by business pursuits, consumer demand, and passionate interest in local food and community action. The primary purpose of the community is to facilitate the production, distribution, and consumption of locally sourced food (within 100-200 miles of the site of purchase). It is unclear when the community initially formed. Agriculture has naturally been an important element of any Denton community since the formation of the city in the mid 1800’s ("History of Denton County," 2011). The community does not feature a clearly defined hierarchy of power or management, and financial profits gained within the community do not trickle up to any one owner. The demand for locally sourced food tends to exceed the supply. Though bonded by common interest, there is no single physical or virtual space where members may access community goods or shared knowledge.

Setting the Stage

Theoretical Frameworks

Increasingly, innovations both nurture and contribute to communities of practice (Lave & Wenger, 1991; Wenger 1998) where individuals learn through their engagement in social practice that is spontaneous, self-organizing, and fluid. It is through participation in informal communities that groups share activities, practices of social communities, construct identities and practice sensemaking (Wenger, 1998). Participation in communities of practice results in exposing and sharing diverse information and knowledge among group members. Lave and Wenger (1991) argue that individuals start out as peripheral members of communities of practice until the newcomer has engaged in learning of knowledgeable skills. Wenger (1998) views communities of practices as a place of negotiation, learning, meaning making, and identity.

Similar to Granovetter’s (1973; 1983) strength of weak ties, communities of practice involve interpersonal connections where the diffusion of cultural ideas develop and disseminate among a network cluster. In the case of innovations or developments, the weak ties are transformed from the periphery and become the center of the network (Granovetter, 1983) to contribute to the community of practice. Nahapiet and Ghoshal (1998) thought that communities were useful narratives for transferring tacit knowledge, exchanging narratives, and preserving rich information. Chiu, Hsu and Wang (2006) found social interaction ties was a strong predictors for developing contribution interest and sustaining knowledge exchange in virtual communities. Digital communities and networks require a critical mass whereby the value of the technology increases as more members are actively participating in that virtual space. The network effect allows for information flow, ease of connectivity, organic growth, and rapid iteration for the improvement of ideas and concepts.

Communities of practice are more influential in looking at how knowledge is created and transferred (Brown & Duguid, 1991; Lave & Wenger, 1991). A growing number of communities of practice use social media and online spaces to share information. Increasingly “individuals voluntarily contribute their time, effort, and knowledge toward the collective benefit” (Wasako & Faraj, 2005). Retna and Ng (2011) found that communities of practice can be an effective mechanism for promoting knowledge transfer both within itself and to other broader organizations. Knowledge is constructed through complex processes of social negotiation and interpretation shaped by the access to resources, reputation, social networks, and access to those involved in developing the knowledge (Greenhow, Robelia, & Hughes, 2009). The social constructivist nature of knowledge itself has not changed (Prawat & Floden, 1994); however the broad impact of social media has increased knowledge fluidity and information sharing.

In a networked society Castells (2004) stated that “power continues to be the fundamental structuring force of its shape and direction…located in the networks that structure society” (p.224). With
the advent of the social web, an increasing number of communities are using this power of the network with online resources that are accessible, easily deployable, and increasingly easy to manage. Zhang & Watts (2008) found that a high level of activity is vital to support engagement between members, which in turn helps members of the community shape the organization. Ardichvili, Page and Wentling (2003) learned that participants were more willing to use the community of practice as a source of knowledge if they trust it as a reliable and objective resource. Online networks make use of social media websites and applications to allow users to customize how they get information, engage in sharing knowledge, and interact within a community based on personal interests or affiliations. The key would be to assess the “virtual settlement” of this online community of practice with regards to the group’s computer-mediated communication (CMC) to understand the interactivity, variety of communicators, sustained membership levels, and the virtual comm.-public space (Jones, 1997).

Organizations are embracing technology and innovation to develop and exchange information at a rapid pace. Argote (1996) indicates that knowledge can depreciate if individuals leave the organization and if technologies become inaccessible or difficult to use. Knowledge transfer is critical for organizational survival and sustainability. By providing communication opportunities and shared documents between members of multiple organizations, transfer of knowledge flows through the general environment and benefit the organizations (Argote, 1996). Virtual networks have the ability to share and distribute information and knowledge among a community of practice. This research paper will explore how knowledge sharing in online communities of practice can improve the organization of this grassroots movement.

Technologies Used

In the scope of this paper, three views of used technology are relevant. First, the technology currently being used by this LFS. Second, the technologies the researchers have arranged, with limited development and financial support, to begin augmenting this LFS into a knowledge sharing social media. Third, the technologies the researchers aim to use in the future.

Through ethnographic fieldwork conducting interviews with community members, observing, and participating in the community both online and offline, researchers identified online presence and use of social media among eight influential organizations in the Denton Local Food System (LFS) Community. A review of online spaces revealed Denton LFS Community use of Facebook Fan Pages (e.g. http://www.facebook.com/pages/Cardos-Farm-Project/230389597022699?fref=ts), Wordpress blogs (e.g. http://www.earthwise-gardens.com/), and moderate use of custom-designed websites (e.g. http://www.kensproduce.com/). Much of the online communication within the community takes place on Facebook and via email newsletters. After speaking with several highly active members of the community, it was confirmed that a large part of the communication takes place via word of mouth or other face to face means.

Currently a number of free web applications are being used to begin organizing the Denton LFS community online. A blog has been arranged and acts as the current version of the researchers’ end product. It serves as a webspace where the researchers can organize and publish the knowledge collected about the community in an accessible manner for community members to access. The applications being used are: 1) Wordpress for the blog www.feeddenton.org, 2) Gmail for email feeddenton@gmail.com, 3) YouTube for video sharing www.youtube.com/user/FeedDenton, 4) Mailchimp for collecting emails and sending newsletters, 5) Ustream for livestreaming LFS events http://www.ustream.tv/channel/feeddenton, 6) Facebook for interfacing with the community http://www.facebook.com/FeedDenton, 7) Twitter for interfacing with the community http://twitter.com/feeddenton, 8) Instagram for photo sharing, 9) HootSuite (http://hootsuite.com/) for managing social media accounts and scheduling posts, and 10) Google Alerts for collecting the latest news on a number of LFS related keywords which are then curated and shared through the social media channels.

Ultimately, the researchers aim to develop a custom knowledge sharing social media that would facilitate user-generated content and cater to the particular communication methods practiced between the various roles in the community, which the researchers have observed through ethnographic methods. The envisioned product will utilize 1) Responsive Web Design to ensure a clean interface for mobile, tablet, and pc users, 2) a Mobile First philosophy, spearheaded by Luke Wroblewski, which advocates designing for small screen mobile devices first in order to push only the most important functionality
forward, 3) the mongoDB non-relational database system for high scalability, and 4) the Node.js server-side javascript language for highly asynchronous, event-driven data, and immediate, real-time updates delivered to users across the network.

Notable Players and Roles

Rarely do members of the LFS community play only one role. The community tends to be filled with passionate members who are willing to be both patrons and community informants. We are aware of several individuals who are producers, vendors, organizers, information disseminators, and patrons. It is individuals like these who comprise the core leaders of the community. Six roles have been listed with brief descriptions:

- **Patrons** - These members purchase and consume locally-grown food and provide monetary compensation to vendors and producers. They are somewhat equivalent to the end user of the LFS.
- **Vendors** - These members serve as intermediaries between producers and patrons. Sometimes will travel to pick up yields, other times will accept delivery from producers. Vendors sell local food to patrons at community markets, shared spaces, or their own storefronts. Some vendors sell raw yields while others, such as restaurants, sell prepared local food.
- **Producers** - These members research and practice agriculture techniques specific to North Texas and are responsible for growing, harvesting and delivering crops to vendors and patrons.
- **Event Organizers** - These members arrange community events around participation and/or education in the LFS community. Some events are recurring, such as the community market, farm to table dinners, and weekly potlucks.
- **Community Informants** - These members bring awareness within and around the community by disseminating LFS knowledge. Often times utilize social media as an outlet for reaching the community.
- **Charitable Participants** - These members donate some kind of service in support of the community. Typically this is in the form of volunteer labor, or free space for events or vendors. For example, one particular bar allows for a local producer to use their back patio as a vendor location every Thursday, and a Montessori school offers a similar service bimonthly for another local food coop.

Purpose of the Study

By utilizing online social networks, the Denton LFS community has the ability to create, share and distribute knowledge in a new and organized way. The purpose of this study is to understand the online environment, motivation for sharing knowledge, and technological needs for this community of practice.

Methodology

Research Questions

1. What community factors create a suitable environment for virtual knowledge sharing community development?
2. Which motivational factors are thought to influence the willingness or tendency of community members to share knowledge?
3. What are the technology requirements to fulfill an unobtrusive ICT design of a knowledge sharing social media for the Denton LFS?
Research Methodology

Our research is in part inspired by results of a pilot quantitative study, which suggested positive willingness on behalf of the LFS community to integrate social media into their infrastructure. A survey was submitted to the community yielding n=68 where all respondents are residents of Denton, TX between the ages of 18-50.

Ethnographic methods have been applied to gather data thus far. The researchers have conducted interviews with members of the community, observed the community, and participated in the community for a number of years. This practice has yielded important regarding member roles, knowledge types, and communication practices that will help to inform the design of an knowledge sharing social media that can best facilitate an unobtrusive transition into a digital communication system.

In the future, a concurrent mixed-methods approach will be applied to allow for researchers to integrate the information when interpreting the results (Creswell, 2009). Ethnographic techniques will allow observation, while the research process remains flexible to the realities encountered in the field to better understand the social and environmental elements of knowledge sharing using social media (LeCompte & Schensul, 1999). This investigation will produce a detailed description of how the existing LFS community operates the modes of community participation, as well as the information sharing needs of the community in order design and implement social media that will support community growth based on knowledge sharing in the LFS setting. Semi-structured interviews, web-based surveys, with open- and closed-questions, and analysis of the current online community documents and artifacts will be implemented to identify system requirements.

Technology Components and Organizational Concerns

Many highly influential producers and vendors of the LFS community are already using Facebook as a primary channel for marketing their product. This combined with the positive technology integration data collected in the pilot study leads us to believe that spending time online to support the community may not be a large obstacle for community members. Preliminary findings from ethnographic research such as interviews, observation, and participation, point to specific functionality requirements for this LFS to effectively share knowledge through social media. Though members have expressed comfort in using such popular sites as Twitter and Facebook, this LFS would benefit from additional functionality not typically associated with these sites, such as archiving and organizing knowledge by type for future retrieval in a highly accessible and unobtrusive manner. LFS knowledge can be divided into four specific categories:

- **Agricultural Practices** - Knowledge generated by and shared amongst producers. Currently stored in individual’s memory and online in a location indifferent manner. Shared most commonly through word of mouth and other face to face modes of communication.
- **Participation Methods** - Knowledge that concerns ways existing members and new members can participate in the community. This may include a list of producers and vendors, volunteer opportunities, or common locations for food pick-ups. Currently stored in individual’s memory as well as lightly around Facebook, Twitter, and Blogs.
- **Spontaneous Event-based Knowledge** - Knowledge concerning time relevant information, much like that which is commonly shared on Twitter. This could be events, sales, or how producers are dealing with inclement climates. Currently shared through word of mouth as well as heavily on Facebook and Twitter.
- **Community Improvement** - Knowledge concerning the state of the community, what is limiting the community, and what could be done to improve the community. This knowledge tends to be generated by the most active members and is commonly shared through face to face communication.

Knowledge will be uploaded by community members in video, photo and also in text formats. Knowledge uploaded will be categorized appropriately by knowledge type and organized accordingly for accessible retrieval. Most of the knowledge has an address and sometimes a date and time attached to them. These
pieces of knowledge will be added to a map and/or calendar. When considering LFS, geographic location tends to be an important factor patrons consider when deciding who to purchase food from.

**Conclusion**

**Solutions and Recommendations**

The Denton LFS features unique member roles as well as knowledge types. Communication patterns within the community are likely to be similar to those of other LFS communities. This LFS shows signs of being willing and ready to organize online. Due to lack of ability to cater to the particular member roles and knowledge types found in this LFS, current popular social networks will not suffice. A custom online application developed in accordance with the unique communication patterns found within this community is necessary to facilitate positive and unobtrusive community growth online.

We believe that a knowledge sharing social media will stimulate community growth and participation, thus increasing social practice amongst community members. Operating under Wenger's axiom of "Engaging in social practice is the fundamental process by which we learn and so become who we are," it follows that we believe a knowledge sharing social media will stimulate learning. Because we aim to design a knowledge sharing social media that is optimized for thoughtful archiving and accessible information retrieval, we believe this application will facilitate lifelong learning by providing a knowledge base the community can rely on and build on top of.

Additionally we suggest that an ethnographic study be conducted over the community of interest. This will aid in providing insight necessary to designing and implementing a social media that is unobtrusive. The social media should fit into the community as seamlessly as possible, and should organize information in a manner that is parallel to how the community already organizes information.

Based on our initial research, we see a need to move forward with the proposed mixed methods study. The opportunity for increased learning and community growth is strong. Members of the community tend to be very receptive to the idea of organizing online and themselves acknowledge the potential benefits. We believe a knowledge sharing social media is an appropriate design to begin facilitating increased informal learning within the LFS community of practice in Denton, Texas.

**References**


Scientific INQuiry (SINQ): Social Media for Everyday Science Learning

Michael Gubbels  
University of Maryland  
College Park  
mgubbels@umd.edu

Jason Yip  
University of Maryland  
College Park  
jasonyip@umd.edu

Jinyoung Kim  
University of Maryland  
College Park  
jkim0204@umd.edu

June Ahn  
University of Maryland  
College Park  
juneahn@umd.edu

Abstract

In this paper we describe SINQ, a prototype mobile social media (SM) application that utilizes social participation to guide learners through an everyday Scientific INQuiry process. The paper outlines the motivation for SINQ based on learning theories of scientific inquiry, the challenges associated with scientific inquiry learning within everyday settings, the design of SINQ to promote science inquiry, and the implications for design and learning with social media that we learned from this development experience.

Keywords: science inquiry, informal learning, social media, scaffolding, natural inquiry

Introduction

The idea of “Science for All” outlines a set of essential thinking skills for citizens of the 21st century (e.g., AAAS, 1990). Opportunities for science are all around us and there are ripe opportunities to learn science in everyday life contexts. Moreover, researchers find that science learning in formal schools is often irrelevant and disconnected from learners' everyday lives (e.g., Basu & Barton, 2007). This gap between real-world contexts and science learning is a major problem that has negative cognitive and affective implications for learners. In order to bridge this gap between formal classrooms and everyday contexts, researchers have undertaken ways to build science inquiry into informal settings, such as museums (Cahill et al., 2011) and afterschool programs (Clegg, Gardner, & Kolodner, 2010). Despite the “fun” experiences that informal settings can provide, science learning is quite complex and requires some of the structured processes found in the formal classroom (Kirschner, Sweller, & Clark, 2006). How do we bridge the structured science learning of classrooms with the engaging experiences children have in everyday life? We believe that social media technologies (SM) can support the ability for learners to engage in natural inquiry as it occurs in everyday life, while also developing formal scientific skill sets through the exploration of personal interests. In this paper, we describe our efforts to develop SINQ (for Scientific INQuiry), a prototype mobile application to support collaborative scientific inquiry activities. We introduce the theoretical framework that underlies SINQ, the design features of SINQ, and the implications of fostering everyday science learning through SM.

Background

Learners come to any learning situation with productive intellectual resources from their everyday experiences that are compatible with scientific thinking (e.g., Hammer & Elby, 2002). For instance, Basu and Barton (2007) observe that learners have “funds of knowledge”, such as lived experiences, cultural practices, and information that is powerful and can set them up to engage in science. One design challenge is figuring out how to leverage new technologies to help individuals use their everyday knowledge to engage with scientific inquiry. Science inquiry practices take the form of


Copyright is held by the authors.
constructing explanations, assessing available sources of information to inform one’s observations, testing hypotheses, and interpreting data or results (NRC, 2000). These practices can be seen in formal, science activities in classrooms (e.g. structured lab experiments), but they are also practices that can provide value in everyday life. For example, whenever someone wonders why the sky is blue or notices that their use of ingredients in their brownie recipe changed its taste, opportunities for inquiry arise that can be lost.

We believe that social media can capture these opportunities to foster lifelong learning in everyday science inquiry processes. SM is a natural way to ask questions and crowdsourcing information. For example, search engines such as SOCL (www.so.cl) utilize social searches and questioning through Facebook™. In addition, social media is accessible and is broadly used across different populations. Increasingly, children (under the age of 13) are becoming major users of social media communities (Grimes & Fields, 2012). Our project begins to address the need for design-based studies that can better understand the design of social media for children’s learning.

**What is SINQ?**

We developed SINQ as a mobile social media application that fosters peer-based science learning across different contexts (Ahn, Gubbels, Kim, & Wu, 2012). In SINQ, multiple learners can contribute smaller pieces of the inquiry process and attach them to others (e.g. questions, hypotheses, project ideas). The system aggregates these contributions into coherent science projects that children can pursue on their own or with friends. SM platforms can be deeply engaging for young people as they build social capital, share personal aspects of their daily lives, and show off their interests through creative means (e.g., Greenhow & Robelia, 2009). However, based on relevant learning theories, engaging in independent inquiry is complex and cognitively taxing (Kirschner et al., 2006). Learners need structure in the process of developing a good scientific question, creating feasible hypotheses for experimentation, and setting up an investigation (Quintana et al., 2004). In our prior work, we designed a basic prototype that could provide scaffolding of inquiry processes through a browser-based social media platform (Ahn et al., 2012). However, through our subsequent work with children we found that learners also need technologies that fit naturally into their daily lives and make science learning personally meaningful (Clegg et al., 2012). Therefore, in this iteration, we have designed SINQ to be specifically mobile. This focus on mobility can encourage learners to use SINQ in their natural inquiry as it occurs in everyday life, while at the same time, helping to scaffold scientific inquiry skills as they explore their own personal and unique interests. This specific iteration of SINQ also focuses on further developing an interface and integrated media tools through co-design with children (Druin, 2002).

SINQ is designed to foster social, collaborative science inquiry in four ways. First SINQ is designed to help learners capture personally meaningful elements of the inquiry process through their everyday interactions with the world. Similar to Pinterest™ we developed SINQ as a way to post questions and capture learners’ daily on-the-go interactions with the world (Cahill et al., 2011). Instead of developing science questions, hypotheses, and investigations in inauthentic settings (e.g., classrooms), SINQ allows learners to enter an inquiry at the moment of inspiration. SINQ allows learners to capture photos of their interest, ask questions or develop a hypothesis based on what is recorded, and post the response to their networks. For example, a child named Alex might come across an observation in which butter does not dissolve in water in his kitchen. Alex uses SINQ on his mobile device to snap a picture of the butter and post his question, “why is the butter not breaking down in water?”

Second, SINQ allows learners to make small and incremental contributions that help to scaffold the process of scientific inquiry. These small contributions from multiple members are then aggregated into coherent science projects. We designed SINQ so that learners can enter into the science inquiry process at any point through contributions of questions, hypotheses, or investigation ideas to match the fluid nature of scientific inquiry. Lastly, the system guides learners through the entire process regardless of where they start. For instance, another child, Barbara, wants to first post a hypothesis about making cake frosting. SINQ prompts her to ask, “what do you notice?” She writes, “my frosting is thicker”. SINQ prompts again, “why do you think it happened?” Barbara thinks and writes, “because it got cooler.”

Third, SINQ provides ways in which learners can participate in social vetting processes through reflection prompts. SM tools, such as Facebook™ and YouTube™ have various voting mechanisms. However, beyond simple voting we believe that the reflection process is crucial for effective learning (Cahill et al., 2011) and is an important aspect in scaffolding in learning tools (e.g., Quintana et al., 2004). Therefore, SINQ is designed with prompts to promote reflections and elicit participation. Another child,
Recommendations for Design and Learning for SM Tools

Based on our experience designing and implementing SINQ we offer the following recommendations for designing social media for fostering informal learning.

Design recommendations

*Use interdisciplinary design.* The challenge of building a social media platform for science learning in everyday life requires collaboration across disciplines. To understand the problem of science learning, we needed knowledge of the current problems in science education and how to approach science learning through frameworks of cognition and the learning sciences. From a computer supported collaborative learning perspective, we needed knowledge of how people approach learning in distributed ways. Finally, to build the SINQ platform, we integrated our knowledge from human-computer interaction, computer science, and information sciences. Approaching this problem from a number of perspectives allowed us to see the challenge of SM learning in science and how to tackle the issue efficiently.

*Co-design with children.* Our belief is that SINQ should be a social media tool that allows learners to contribute questions that inspire them from their everyday experiences. Using SINQ must feel natural, simple, and easy. Therefore, in order to better understand what learners would want in SINQ, we co-designed the technology with child partners at the University of Maryland’s Human-Computer Interaction Lab. Using Cooperative Inquiry (Druin, 2002) we worked together with children (ages 7–11) to design SINQ in a way that would be usable and feasible in an everyday setting. We asked children to develop low-tech prototypes of SINQ and how they wanted to pose questions and hypotheses. Our design sessions with the children gave us insights into how to build an interface that is both structured enough to scaffold science yet inviting to children, including interface design suggestions, how to elicit meaningful science questions through natural interactions, and how to design the voting mechanisms.

*Developing with a learning environment.* We also worked with an informal science education program called Kitchen Chemistry (KC) (Clegg et al., 2010). In KC, children engage in the design of their own food investigations to learn more about the science inquiry process. We have observed in KC that when learners made contributions to SINQ and searched through the contributions of others, they could 1) find other learners with similar inquiry interests, 2) help each other through suggestions and refinements, and 3) develop new questions, hypotheses, and investigations based on other postings. These qualitative findings helped us confirm the efficacy of using social media in the context of scientific inquiry (Clegg et al., submitted).

Everyday learning and Social Media recommendations

*Strive for natural and mobile interaction design.* Science inquiry in everyday settings means that social media tools need to allow learners to make contributions quickly, simply, and ubiquitously from learners’ interactions in their natural environment. For SM to become prevalent in everyday learning in informal environments, the tools must be able cross borders between home and school. In our prior work, we developed a system that was primarily browser-based (Ahn et al., 2012). However, in our subsequent co-design work with children, we quickly found that this mode of interaction was not natural or engaging for a child population. The children co-designers provided us with deeper insights and experiences about the importance of mobile and ubiquitous modes of learning with technology.

*Distribute complex inquiry processes.* Science inquiry processes are challenging and complex. We believe social media can distribute the cognitive demands of science inquiry so that learners can
engage in science practices more easily. Science inquiry is a social process; no single person can truly understand the entire inquiry process in isolation. Evidence-explanation frameworks of science inquiry call for more conversations and examinations of dialectical exchanges between observations, data, and theory. Duschl (2008) observes that such conversations include argumentation, debates, modeling, drawing, writing, and other forms of communication. Through social media children across or within settings can select data for evidence, make claims about evidence, create propositions, and evaluate explanations. Therefore, we suggest that the small contributions of each individual can be assembled to create a crowdsourced inquiry process that allows learners to engage in science on their own personal terms.

Provide natural scaffolding through SM feedback mechanisms. Social vetting such as voting and ratings can act as natural scaffolds to help participants learn what makes a good question, a testable hypothesis, or a feasible experiment. We found that guidance from a teacher or facilitator is helpful, but that learners own social interactions in SINQ allowed for more genuine collaborations and refinements. Through learners’ own reflections and examination of other contributions, we believe learners are able to develop their own collective strategies to engage in science inquiry.

References


iMapLibraries: Mapping Opportunities for Lifelong Learning

Laura Spears  
Florida State University  
lib03@my.fsu.edu

Nathaniel Ramos Jr.  
Florida State University  
nr12c@my.fsu.edu

Jisue Lee  
Florida State University  
jl10n@my.fsu.edu

Abstract

The iMapLibraries project combines social media with online mapping to help users find free lifelong learning classes available at local public libraries. Data on classes is solicited through a social media campaign which invites public libraries to “put themselves on the map.” Libraries can add data about classes directly into the iMapLibraries tool through an online registration page. Data is collected into a Google Fusion Table which produces a Google Map of the libraries with five categories of classes offered plus clickable links for more information via the library's web site, Facebook, Twitter, and calendar page. iMapLibraries offers potential for improving public access to lifelong learning classes in libraries, and demonstrates new ways for libraries to collect and share data on available lifelong learning programs and services through social media and online mapping. Future possibilities include mapping other library use data such as meeting spaces, book groups, and tutoring services offered in different languages, which could be found on interactive maps accessible on the web, through social media, and via mobile apps.

Keywords: libraries, social media, mapping, GIS

Introduction

Libraries continue to provide a traditional service to communities in that they bring people and information together (Prentice, 2011). Although this purpose has shifted in response to user adoption of personal technologies and the proliferation of household Internet access, the library's role as the intermediary for those who are challenged by digital technologies’ access barriers such as digital literacy, expensive devices and inconsistent or costly service provision, continues to grow. Public libraries have embraced the role of information technology facilitators, providing computer access, training, digital collections and virtual entertainment, and supporting lifelong learning. According to the Public Library Funding and Technology Access Study, 2011-2012 (Hoffman, Bertot & Davis), libraries are the “lifeline to the technology resources and digital skills essential to full participation in civic life and in the nation’s economy. Libraries continue to transform lives by providing critical services and innovative solutions to technology access” (2012, p. 6). Over 96% of public libraries reported that they provide assistance to patrons seeking e-government services in the form of computer and Internet access, website navigation and even in completing applications for government financial assistance (Hoffman, Bertot & Davis, 2012). However, an ongoing challenge for libraries is to make visible to local learners the wide variety of programs, services and resources for lifelong learning available for free to the community in local neighborhood libraries.

The iMapLibraries project explores social media and online interactive mapping as a way to connect people to libraries, and to connect libraries to each other in building a social media-based research community of practice. Libraries use social media to connect with the iMapLibraries project, and interactively "put themselves on the map" by sharing data on their programs and services.

Acknowledgements: This research is funded in part by the Institute for Museum and Library Studies in a 2012 National Leadership Grant led by Christie Koontz, Dean Jue, and Lorri Mon of Florida State University.


Copyright is held by the authors.
Phase 1: Exploratory Research

The first phase of the project involved exploratory work on developing a coding scheme for the lifelong learning classes, and testing potential online mapping tools.

Lifelong Learning Classes

To identify categories of free classes offered at local libraries that support adult lifelong learning in everyday life, a small-scale pilot study was conducted by searching the web sites of 30 public libraries nationwide. Five major types of lifelong learning classes were identified:

- **Computer** classes included introductions to basic software such as Windows, Microsoft Office applications, Internet searching, email basics, and social media, as well as how to use e-books, iPad and Kindle e-readers, and instruction on avoiding online identity theft and scams.
- **ESL** classes included English language tutoring programs, including providing conversation partners and regular lectures.
- **Government** related classes included instruction on social security disability and supplemental security income, social security, food stamps, tax refund assistance and citizenship testing.
- **Health** related classes featured Medicare and health insurance, and talks on breast cancer, strokes, cardiovascular diseases and more.
- **Workforce** readiness classes included resume writing workshops, interview skills, job-searching and seeking skills, and GED testing assistance and tutoring.

The coding scheme for these five major types of lifelong learning classes offered at public libraries [Computer, ESL, Government, Health, and Workforce] formed the basis for subsequent data entry and data collection.

Social Media for iMapLibraries

Concurrently with the exploratory phase for identifying class types, social media networking was launched, setting up the social media sites to be used in the iMapLibraries project with a Wordpress blog, Facebook, Twitter, and YouTube. An iMapLibraries web site was also launched (http://www.imaplibraries.org). From September through December 2012, the iMapLibraries social media built a network of social media connections with U.S. public libraries and library community leaders through a multifaceted Facebook, Twitter, and blogging campaign, and disseminated information about the iMapLibraries mapping and data collection project.

The iMapLibraries social media network also sought to connect with opinion leaders in the library community. Connecting with key influencers or information “gatekeepers” (Metoyer-Duran, 1993) can influence wider dissemination of information through trusted leaders in a community.
Exploratory Mapping

Initial testing of mapping tools for the project involved using Google Maps by inputting library data on lifelong learning classes, calendar links and social media sites into a shared Google Map (see: https://maps.google.com/maps/ms?msa=0&msid=207581887152590798604.0004cc2d2bd4866020372). An ideal mapping tool for the project would allow participating libraries nationwide to easily “add themselves to the map,” interactively adding their own data on their library and their lifelong learning classes for learners to be able to find on the map. From these exploratory efforts, it was determined that key requirements for a mapping tool included 1) support for shared data entry by many participants, 2) high usability, in terms of being easy and quick to use, and, 3) potential for usability even at a low level of technical skill, in order to facilitate wide-scale use among a diverse range of participants.

Ultimately, Google Maps was found not to be feasible for iMapLibraries due to a slow data entry process, and the need for participants to have HTML knowledge to use the tool most effectively. Google Fusion Tables emerged as an alternative possibility for shared participatory mapping (see: https://www.google.com/fusiontables/DataSource?docid=1HLwfV6GRjAlcY6a7R-XqvuHFon8cybw--2cGAGM#map:id=3). Using the five coding categories for classes, a Google Fusion Table was created that could receive data typed in by library participants through an embedded Web-based registration form, generating an online map of libraries with lifelong learning classes.

The Google Fusion Tables allowed a distributed and participatory approach in which participants at libraries nationwide entered their data into the map via a simple Web-based form that did not require HTML knowledge or other special skills. Attempts were also made to write a script that would update and refresh the resulting Google Map periodically to reflect new data entered through the online Google Fusion data collection form. The Google Fusion Tables map was embedded on the iMapLibraries site (see: http://imaplibraries.org/maps.html), allowing libraries to participate by directly entering their own data on lifelong learning classes (see: http://imaplibraries.org/register.html).

Phase 2: Project Launch with Google Fusion Tables

In further testing of the Google Fusion Tables functionality for data collection and mapping, possibilities of filtering by variables were tested. Using Census Bureau 2010 diversity data on libraries serving diverse communities within 2-mile radius of the library branch, an additional Google Fusion map showing public libraries serving categories of diversity was created (see Figure 2).

Figure 2: Google Fusion Tables Map for iMapLibraries: Libraries Serving Diverse Communities
https://www.google.com/fusiontables/DataSource?docid=1BnT2Dugkot2tJv69P6vdYZ7AaGm7b4MFg_Xq8ZY
Different colored and shaped markers in the Google Fusion Tables map were used to indicate the range of diversity served by the library within the local community population, as follows:

- Red Push Pins indicate libraries with 15+ diversity variables within two miles of the library
- Yellow Push Pins indicate libraries with 10-14 diversity variables within two miles of the library
- Red Dots indicate libraries with 5-9 diversity variables within two miles of the library
- Yellow Dots indicate libraries with 1-4 diversity variables within two miles of the library

In filtering for the different variables, public libraries using the Google Fusion Tables map could locate other libraries serving similar patterns of diversity variables within their local populations to share and compare programs and services offered for learners and contact the other libraries from map web and social media links or through the iMapLibraries social media network. This experiment also demonstrated the feasibility of filtering by other variables in Google Fusion Tables, such as users filtering the lifelong learning classes map to find libraries that offered only specific types of classes.

**Conclusion**

iMapLibraries explored potential tools for connecting public libraries with online mapping via social media, finding that Google Fusion Tables offered useful functionalities for participatory mapping and filtering by variables. The experimental mapping project resulted in three interactive maps: 1) a Google Maps-based map of lifelong learning classes at U.S. public libraries; 2) a Google Fusion Tables-based map of lifelong learning classes at U.S. public libraries, and 3) a Google Fusion Tables-based map of public libraries serving diverse local populations with a 2-mile radius of the library’s location, using diversity data drawn from the U.S. Census.

In its potential for mapping libraries and their services offered, the iMapLibraries project helps learners to find lifelong learning opportunities in libraries, and by also mapping the diverse demographic and socioeconomic populations that libraries serve, iMapLibraries also helps librarians to learn more about their community and what local needs are for lifelong learning resources, programs and services. Social media connected libraries to iMapLibraries and to each other, and embedded interactive mapping into the social network discourse. Social media also offered potential for bringing maps to potential learners via Twitter, Facebook, blogs, as well as news feeds and media outlets. In using social media with interactive mapping, iMapLibraries could make more visible to local learners the wide variety of programs, services and resources for lifelong learning that are available for free in local neighborhood libraries.

Ongoing implementation efforts by iMapLibraries will continue to explore how participatory online mapping can be used to collect and share research data about library programs and services. Future possibilities include mapping other library use data such as services offered in different languages, and developing mobile apps for user populations who may not have access to computers and Internet, but may be able to access information using mobile technologies.

**References**


#ElectionClass: Encouraging Strategic Thinking About Politics Through Social Media

Jared Mandel
Syracuse iSchool
jsmand02@syr.edu

Chelsea Orcutt
Syracuse iSchool
cдорcutt@syu.edu

Abstract

#ElectionClass is a course on social media and the 2012 election offered by the School of Information Studies at Syracuse University. The course involves in-class tweeting, analysis of the real presidential election, as well as a mock campaign project. A number of students took on additional interdisciplinary efforts that emerged as a result of the coursework. #ElectionClass's experiential teaching model promotes students' lifelong learning and engagement with social media by giving them a voice and a role in a real-life event.

Keywords: social media, interactive, data, learning, politics

Introduction

The phrase was featured all over the news throughout the past few months: the 2012 election was the most social election of all time. The millions of tweets, Facebook likes, and Instagram shares that were generated by the Obama and Romney campaigns were a manifestation of the success of true voter engagement.

The American people had the opportunity to be involved in the political process on a whole new level, thanks to the campaigns' use of social media. This was many citizens' first exposure to social networks, and their experience with social media has motivated them to continue using these networks long after the election has ended. In short, the social media use during the election was a catalyst for what could be a lifelong desire to learn.

A group of approximately 120 Syracuse University students had the chance to experience the 2012 election in an even more revolutionary way. Enter: #ElectionClass.

What is #ElectionClass?

IST 500: Social Media and the 2012 Election, more commonly known as #ElectionClass, is an immersive experience that challenges students to think strategically and develop detailed political campaigns rooted in social media.

The course, designed and taught by Professor Anthony Rotolo, aims to explore the impact of social media on elections, political movements and citizen engagement with a special focus on the 2012 American Election as it unfolds in real time. Throughout the course, undergraduate and graduate students from all majors across the University use the real election as the lens for viewing how emerging technologies and citizens' social media engagement impact candidates, campaigns, voters, and success at the polls.

Acknowledgement: NEXIS social media research lab; special thanks to Dean Liz Liddy, Anthony Rotolo, Chris Becker, Andrew Bauer, Jared Kraham, Diane Stirling, Hailey Temple.
Copyright is held by the authors.
Course Structure

During class, students engage in lectures by live-tweeting their thoughts and responses (tagged with #ElectionClass), which are displayed on a Twitter feed at the front of the room. The #ElectionClass hashtag has nationally trended on Twitter nearly every week throughout the semester. On average, over 1,300 tweets using the #ElectionClass hashtag are generated in each three-hour class.

Students’ careful analysis in class often results in extremely interesting findings. But the major #ElectionClass project challenged students to do much more than observe the American political campaigns. Instead, they organized their own.

The #ElectionClass Campaign

At the beginning of the course, undergraduate students were randomly divided into 14 campaign teams. Each six-to-eight person team was assigned a campaign candidate persona. Personas ranged from "democratic senator from New York State," to a “republican governor from Alaska.” These campaigns would take place in "Amercia," an alternate universe and America’s “sister country,” whose name is inspired by a spelling error in a mobile application launched by the Romney campaign. Meanwhile, the graduate students in the course took on the role of members of the media that would cover the campaigns as they unfolded. The course syllabus detailed the campaign project, weekly reading assignments, blogging assignments, and quizzes.

On the first day of class, each team selected a student to play the role of their candidate, and the race was on. By the end of class that day, campaign Twitter accounts were launched, website domain names were purchased, and campaign strategies were formulated. The goal? Garner enough support throughout the course to win the in-class Amercian election, held on November 6, just like it was in the real world.

Learning Goals

Each candidate and team was tasked with executing a complete social media strategy, leveraging any networks they wished. Nearly every team created a Twitter and Facebook page in addition to realistic-looking websites. Tweets and Facebook posts included everything from information on the candidates’ stances on foreign and domestic issues (which mirrored those of America), or interactions with their “constituents” (classmates) and fellow candidates.

The student candidates also participated in two primary debates, moderated by Professor Rotolo, which brought with them all of the excitement, controversy, and social media frenzy that was seen in the Obama-Romney debates. Candidates were eliminated based on live straw poll results after each debate, until one republican and one democrat were selected.

Campaign Progression

In true #ElectionClass fashion, the unexpected happened. An independent candidate, Carter Rhodes (played by iSchool student Brad Slavin) emerged to join the Democrat, Matt Diaz (played by Jared Kraham), and Republican, Robert Lawrence (played by iSchool student Bob O’Brien). Rhodes created his own political party (the “Everyone Is Invited To This [Party]”) and after securing enough signatures on a petition, was permitted to participate in the final debate.

The emergence of an independent candidate was not the only deviation from the course syllabus. Because of their excitement about social media and their desire to use their skills in innovative ways, a number of students developed projects that were incorporated into the class. The students that were most successful in the course demonstrated a genuine desire to apply their skills outside of class and go above and beyond the required projects. These students tweeted relevant articles and insights not only during the three-hour class each week, but throughout the week. They religiously followed news outlets and watched the presidential debates. Students who failed to track the news and put their coursework into practice may have struggled to find the true value in the course.
Additional Interdisciplinary Efforts

Because the focal point of #ElectionClass is the mock campaigns, it quickly became clear that a fair Election Day voting method would need to be developed, especially because final voting was not only open to #ElectionClass members, but also to the entire Syracuse University campus.

#ElectionClass Voting System

In keeping with the trend of students going above and beyond the syllabus requirements, iSchool seniors and #ElectionClass students Andrew Bauer and Chris Becker developed a revolutionary voting dashboard using a full LAMP (Linux, Apache, MySQL, and PHP) stack that allowed students to cast their votes on a touch screen, with the results being tallied in real time.

The dashboard contained a list of the candidates and their political stances, as well as boxes for students to click as they cast their votes. After the initial vote was cast, students were led to a new screen and asked to answer exit poll questions related to each candidate’s use of social media, quality of political ads, and debate performance. Bauer and Becker compiled data from exit poll questions to provide the class with real-time updates throughout Election Day.

#ElectionClass Voting Results

Voter feedback indicated that the #ElectionClass dashboard provided a more favorable voting experience than the paper ballots used for the real election. Over 500 votes were cast in the mock election, which exceeded the number of real election votes cast at the on-campus local polling place. This was a clear manifestation of the success of #ElectionClass students’ passion for putting what they’ve learned into action.

Figure 1
#ElectionClass Final Presidential Voting Results

Collected November 6, 2012

This chart displays data collected from 9 AM to 7 PM on November 6, 2012 in the NEXIS social media research lab.

What #ElectionClass Illustrates

The interactive nature of #ElectionClass demonstrates that when students are passionate about a topic that is timely and relevant to their daily lives, they will go far beyond what is required of them, simply because they’re excited about it. The strong parallels between the course and the real election allowed students to feel a sense of accomplishment as they experienced political success.
The nature of the class, particularly the presence of the Twitter stream, encourages students to make their voices heard across social media. The campaign teams thrived by thinking critically about their social media strategies, and were motivated by a desire to spread their messages and encourage their constituents to show their support on social media, and ultimately, with a vote. The candidates’ reputations were on the line, and each team worked hard to preserve them.

The course syllabus did not require campaign teams to complete specific social media tasks. For example, there was no assignment for campaigns to send out a certain number of tweets, or to obtain a minimum number of Facebook likes. The campaign project turned students into forward-thinking self-starters, as they crafted their social media strategies based on what they were learning, not what they were forced to do. Students recognized that there was more than just their grade on the line; they also had to protect their candidate and team’s reputation. They were motivated to do additional work and create a myriad of extra campaign materials and social media accounts, simply because the course encouraged friendly competition.

The voting dashboard project was born out of students’ desire to build something better than what is currently available. The final product was created using skills the students developed in the iSchool, and required a thorough understanding of social media engagement. Moreover, the dashboard technology can now be used in future elections, on campus and beyond. It’s a physical contribution from a course rooted in digital.

**Conclusion**

The legacy of #ElectionClass lives on not only through the voting dashboard, but also through the campaigns themselves. Campaign teams created stickers, buttons, and posters to supplement their social media strategy, and those materials can still be found all over campus. The media coverage the course received, from local news and national media, including the Today Show, intrigued members of the campus community and encouraged them to see the role of social media in a new light.

#ElectionClass demonstrates that social media provide a way to foster a unique learning experience. The innovative teaching method employed by the course provides a model for future courses based on interactivity, timeliness of the course topic, and encouraging students to compete to develop the best campaign or project. It is an effective vehicle to engage students to make their voices heard, and empower them to search for real-life applications of social media that will allow us to experience elections and other events like never before.

**Reference**

Analyzing How Users Utilize ‘Riff’ for Collaborative Searching and Sharing Contents for Social Learning in So.cl

Erik Choi
erikchoi@gmail.com
Vanessa Kitzie
vkitzie@gmail.com
Chathra Hendahewa
chathra.hendahewa@gmail.com
Rutgers, The State University of New Jersey

Introduction

Microsoft So.cl combines the collaborative features of a social network and the capabilities of a search engine to create a unique social search environment. The activity of riffing provides a unique twist on a typical search results page since users can actually contribute to the results of another person’s search string, and/or annotate their own searches, also providing multimodal content such as image, video, and text within one results display; as well as user generated metadata via comments and tags. These collaborative and search features give so.cl the potential to promote discovery through interaction with search result content. Unfortunately not many so.cl users exhibit riffing behavior. For this reason, we explored behaviors of users who had riffed in order to further examine their behaviors, focusing also on some of the top riff posters, to determine both what behaviors and post activity might be popular among these users. Results from this analysis indicate that overwhelmingly users who riff on others’ posts also tend to create their own riff postings, as well as the fact that as users tend to become more acclimated to the so.cl service, they also tend to riff more. These findings can be used to determine implications for pushing certain behaviors within so.cl to promote riffing behavior.

Riffing

Of the several features that Microsoft So.cl provides for sharing and organizing content, our specific focus is on “riffing” behavior. Riffing behavior differentiates Microsoft So.cl from other social network sites such as Facebook, Twitter, or Pinterest. When you “riff” on a user’s post, you have the ability to re-post a web object or piece of information that someone has shared and add your own annotations including comments and tags. People who look at your riff have the ability to see all of the riffs emanating from the original post thereby exposing them to research that has already been conducted on their topic of interest.

Incorporating riffed content into a results display not only allows people to be serendipitously exposed to user curated content, as well as pages ranked via a conventional web algorithm, but also allows users to search different content (e.g. text, images, gifs, video, audio) within the same results page. This differs from other commercial search engines in which the user must determine the type of content that they would like to search (e.g. video, images, text). By predicting the content and type of information that users are likely to riff, we can make some inferences about how well two users with similar topical interests can benefit from collaborative learning such as following each other, sharing content and by riffing each other’s posts.

Research Questions

The study focused on determining how riffing relates to other behaviors, specifically providing feedback on riffed posts and posting other online content. Specifically, we would like to explore the following research questions:

Copyright is held by the authors.
RQ1: What are the textual and non-textual features that contribute to a specific piece of content getting riffed?
RQ2: Who are people who frequently riff? How are they the same or different from other users who may not riff as much?
RQ3: How does so.cl activity relate to riffing behavior over time?

Data Analysis

The data was extracted from so.cl Behavior Data files, which were provided by Microsoft, with a total number of 313,745 unique users. Out of this user base, we extracted 4,042 unique user ids of people who have exhibited riffing behavior, or 1.29% of the available data. From here, the study focused on the riffed posts a user created, other riffed posts the user might have provided some form of feedback (e.g. link, comment) on, as well as any other content that the user posted within the so.cl environment.

From a random sample of N=1,000 UserIDs, the study found that the number of other online posts outside of riffing (n=582,598) greatly outnumbered riffing content (riffed posts, n=12,079; feedback on riffed content=11,405) among users who had displayed riffing behavior. The most popular type of online behavior is posting a photo (n=246,684), which accounts for 40.7% of the total content measured - riffed posts, feedback on riffed content, and other posts (See Table 1).

Table 1. Descriptive Statistics on behaviors of users who riff

<table>
<thead>
<tr>
<th></th>
<th>Total Riffs</th>
<th>Own Riffs</th>
<th>Riffs to Others</th>
<th>Total Feedback Riffs</th>
<th>Own Feedback</th>
<th>Riff Feedback to Others</th>
<th>Total Other Online Posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>N = valid 1,000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Std. Deviation</th>
<th>Variance</th>
<th>Skewness</th>
<th>Std. Error of Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12.08</td>
<td>1.00</td>
<td>1.00</td>
<td>58.854</td>
<td>3463.75</td>
<td>8.67</td>
<td>.077</td>
</tr>
<tr>
<td></td>
<td>10.21</td>
<td>1.00</td>
<td>0.00</td>
<td>49.828</td>
<td>2482.78</td>
<td>8.60</td>
<td>.077</td>
</tr>
<tr>
<td></td>
<td>1.87</td>
<td>1.00</td>
<td>0.00</td>
<td>10.967</td>
<td>120.27</td>
<td>9.20</td>
<td>.077</td>
</tr>
<tr>
<td></td>
<td>11.41</td>
<td>1.00</td>
<td>1.00</td>
<td>55.123</td>
<td>3038.52</td>
<td>8.87</td>
<td>.077</td>
</tr>
<tr>
<td></td>
<td>1.69</td>
<td>0.00</td>
<td>0.00</td>
<td>9.090</td>
<td>82.63</td>
<td>9.81</td>
<td>.077</td>
</tr>
<tr>
<td></td>
<td>9.71</td>
<td>1.00</td>
<td>1.00</td>
<td>48.226</td>
<td>2325.78</td>
<td>9.14</td>
<td>.077</td>
</tr>
<tr>
<td></td>
<td>582.60</td>
<td>31.00</td>
<td>10.04</td>
<td>2508.590</td>
<td>6293025.38</td>
<td>137.46</td>
<td>.077</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Kurtosis</th>
<th>Std. Error of Kurtosis</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>88.99</td>
<td>.155</td>
<td>1</td>
<td>793</td>
<td>12079</td>
</tr>
<tr>
<td></td>
<td>85.56</td>
<td>.155</td>
<td>0</td>
<td>653</td>
<td>10207</td>
</tr>
<tr>
<td></td>
<td>96.75</td>
<td>.155</td>
<td>0</td>
<td>163</td>
<td>1872</td>
</tr>
<tr>
<td></td>
<td>94.49</td>
<td>.155</td>
<td>0</td>
<td>764</td>
<td>11405</td>
</tr>
<tr>
<td></td>
<td>94.49</td>
<td>.155</td>
<td>145</td>
<td>445</td>
<td>1691</td>
</tr>
<tr>
<td></td>
<td>10.90</td>
<td>.155</td>
<td>0</td>
<td>681</td>
<td>9714</td>
</tr>
<tr>
<td></td>
<td>137.46</td>
<td>.155</td>
<td>681</td>
<td>45963</td>
<td>582598</td>
</tr>
</tbody>
</table>

A linear regression was performed using number of riffed posts as the dependent variable in order to better examine what types of content (feedback on own riffed content, feedback on others' riffed content, other post types - photo, note, link, comment, tag, video, news, document) were making the most significant contribution to explaining the variability of total number of riffed posts. Surprisingly, given its low sample size (n=9,714) when compared to other online content, feedback on another user’s riffed content had eight or more times the beta value size of any other factor included in the regression (β=0.861). A subsequent stepwise regression revealed that feedback on another user’s riffed content explained 98.2% of the variability of number of riffed posts (r=0.991). On the other hand, photo posts constituted one of the only variables that did not make a significant contribution to the model (p=0.477). This is particularly surprising given that the large sample size of photo posts alone would suggest statistical significance (See Table 2).
Table 2. Result of linear regression on number of riffed content

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Unstandardized</th>
<th>Standardized</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>-.055</td>
<td>.173</td>
<td>-3.21</td>
<td>.748</td>
</tr>
<tr>
<td>Feedback on Own Riffed Content</td>
<td>.864</td>
<td>.036</td>
<td>.133</td>
<td>.2419</td>
</tr>
<tr>
<td>Feedback on Other Riffed Content</td>
<td>1.051</td>
<td>.007</td>
<td>.861</td>
<td>143.647</td>
</tr>
<tr>
<td>OtherPostPostType_Note</td>
<td>0.20</td>
<td>.004</td>
<td>.046</td>
<td>4.836</td>
</tr>
<tr>
<td>OtherPostPostType_Photo</td>
<td>.000</td>
<td>.000</td>
<td>-.004</td>
<td>-.712</td>
</tr>
<tr>
<td>OtherPostPostType_Link</td>
<td>-.011</td>
<td>.003</td>
<td>-.057</td>
<td>-3.746</td>
</tr>
<tr>
<td>OtherPostPostType_Comment</td>
<td>-.051</td>
<td>.006</td>
<td>-.154</td>
<td>-8.760</td>
</tr>
<tr>
<td>OtherPostPostType_Tag</td>
<td>0.21</td>
<td>.002</td>
<td>.068</td>
<td>10.139</td>
</tr>
<tr>
<td>OP_Like</td>
<td>0.10</td>
<td>.002</td>
<td>.075</td>
<td>4.490</td>
</tr>
<tr>
<td>OP_Video</td>
<td>-.33</td>
<td>.005</td>
<td>-.095</td>
<td>-7.225</td>
</tr>
<tr>
<td>OP_VideoParty</td>
<td>-.074</td>
<td>.047</td>
<td>-.005</td>
<td>-1.575</td>
</tr>
<tr>
<td>OP_News</td>
<td>.072</td>
<td>.009</td>
<td>.075</td>
<td>8.386</td>
</tr>
<tr>
<td>OP_Like-Comment</td>
<td>.055</td>
<td>.004</td>
<td>.093</td>
<td>14.071</td>
</tr>
<tr>
<td>OP_Doc</td>
<td>-.750</td>
<td>.346</td>
<td>-.006</td>
<td>-2.169</td>
</tr>
</tbody>
</table>

The variances for the total number of riffed posts, feedback on riffed content, and other online posts are similar relative to the sample size and positively skewed, which indicates that a smaller number of users provide the majority of the content. For this reason, the study decided to further examine top users and their riffing behaviors. The top 25 users across the following categories were first focused: total number of riffed posts, total number of riffed feedback, and total number of other online posts. The study found a 92% agreement between the top 25 users among riffed posts and feedback on riffed posts, and ~60% agreement between top users of these categories and top online posters. Across each category, users also posted ~60% of the total content. From here, top 25 posters of riffed content were focused to compare their cumulative amount of so.cl activity to the other 9,975 sampled users. Independent t-tests assuming non-homogeneous variances indicated that there are a statistically significant differences between the amount of so.cl activity among the top 25 posters and the other users (p<0.01) for all areas except for amount of postings of documents (this could be explained by the relatively small sample size, n=17).

Finally, the study attempted to examine how the activity of our top users has changed over time. Did users tend to create riffed posts more, less, or the same amount after they became acclimated to the so.cl site? To measure this, the study first calculated the median time and date of online postings for each user; the median was chosen as an indicator of acclimatization. Then, the study also counted the amount of riffed posts they created before and after this time period. Among our top 25 users, findings indicate that there was a 66% increase of riffed posts (n_{before}=2,981; n_{after}=4,954) created post-acclimatization to interacting within the so.cl environment, as indicated by online posts. This suggests that among top users, the more time they spend on the site, the more they create riffed posts. Further study could be performed to determine when rate of riffed posts created plateaus and how this might correlate to other post activity within so.cl.

Discussion

One key finding here is the link between a user creating a riffed post and providing some sort of feedback (e.g. comment, link) on another users’ post. This suggests that promoting both activities will marry the objectives of so.cl to both push the social interaction aspect via riffing on others’ posts, as well as the customization aspect of interacting with a search results page via riffing. Future studies should be performed in order to further explicate this link and perhaps measure the effect of interventions meant to enhance these behaviors.
Another finding examined the types of users who frequently riffed. We see that their so.cl behaviors as a whole, both riffing and creating other online posts, are significantly different from the rest of the so.cl users who have riffed. Over time, the study also noticed that the frequency of these behaviors tends to increase. These findings are packed with ideas for future study, including a longitudinal analysis of riffing use over time, as well as how behaviors of top users might differ from other users.

A limitation of this study was that it did not focus on people who did not create riffed content. Future studies could use information about these people in order to create a logistic regression model that could indicate what behaviors, user characteristics, and textual/non-textual features of content shared, contribute to a person’s likelihood to riff.

Conclusion

There is a hope that findings will inform the creation of a push system that attempts to diffuse information across networks by suggesting specific content to be riffed to other users. This kind of information suggestion is important, since it can lead to discovery of new information that is more likely to be of interest to a related user that has common interests with the person who riffed the content, but also brings a certain amount of uniqueness to the content based on the looser social relation to someone who has the potential to be outside of the user’s core network.

Riffing provides a unique way of which the users are able to interact with each other based on their specific interests within Microsoft So.cl network. At this instance, investigating user behaviors related to riffing constitutes the first step for gaining a better understanding of how So.cl users develop social relationships and collaborate. We hope that findings will inform what kind of content is riffed and encourage users to collaboratively search and share content by developing rich user interactions which would pave the way in promoting a mutually beneficial social learning paradigm.
Collaborative Timeline:
Toward the Next Generation of History Learning

Yanqi Wu
kenwyq@umich.edu

Jin Yan
mjinyan@umich.edu

Yanting Zhang
yantingz@umich.edu

Xiaohan Yu
seanyu@umich.edu

University of Michigan

Abstract

In order to learn history, it ought to be a lifelong learning experience, an engaging one, we strive to build a site that engages people into history and helps them navigate across history with advanced technology. A Collaborative Timeline helps people to be more engaged in history learning through effectively managing and sharing history knowledge with friends and fostering collaboration and in-depth discussion among timeline editors. Our timeline makes history fun to learn and easy to illustrate, allows references between historical events and news in present time, enables users to view historical events through different perspectives and facilitates community formation and collaborative learning around history topics. The site achieves these goals by presenting the content using multimedia, offering an intuitive user interface, aggregating different viewpoints and opinions of historical events, enabling a wide range of social features and gamified evaluation mechanisms.

Keywords: collaborative learning, history learning, timelines

Introduction

“Study the past if you would define the future,” as Confucius, a Chinese thinker and philosopher, remarked 3000 years ago. A curious nature and the fear of uncertainty drives humans to get a better understanding of why things happen, and what things will happen throughout their own life. Therefore, studying the past and its relations with the future destined to be a lifelong learning process. However, from interviews with potential users, we found that a significant proportion of users consider the process of learning history somewhat boring, neither fun nor vivid at all. Through their description of their individual learning process, there is a breakdown in users’ workflow of reading current news and referring to historical events. In addition, these input are scattered; therefore, users are in need of a better organizing tool to manage their pool of history knowledge and electronic artifacts. Furthermore, though users are intrigued by the hidden connections and different perspectives on historical events, they are not equipped with a tool to do this.

Motivation

We have conducted user research with 5 users before moving into the system design phase. The potential users we interviewed are from different age groups with different professional backgrounds. What they share in common are their interests in history and eagerness to learn history in a more effective way. From those interviews, we found that a significant proportion of the users consider the


Copyright is held by the authors.
process of learning history is somewhat boring and inefficient. Through their descriptions of their individual learning process, there are difficulties and obstacles in understanding the bigger picture of a specific history period and recalling historical facts. In addition, these input are scattered; therefore users are in need of a more engaging way to learn history facts, a more intuitive method to explore history knowledge, and a better organizing tool to manage their pool of history knowledge and electronic artifacts. Furthermore, though the hidden connections and different perspectives on historical events intrigue users a great deal, they are not equipped with a tool to do this. Our system functions will fulfill their needs of an intuitive, immersive, and novel way to learn history.

Related Work

Current history learning resources are primarily individual, and searching-based. We carefully examined two representatives of them: History.com and Wikipedia.

**History.com:** An educational website contains shows, videos, news, topics and games resources related to history learning. It originated from a cable TV channel formerly known as The History Channel, which broadcasts a variety of fictional and non-fictional historical content, together with speculation about the future. This site uses multimedia to present history knowledge in a vivid way. However, the information is organized poorly and supports passive learning only. In contrast, our system supports timeline folder as a powerful tool to help users manage their history knowledge and utilizes the advantages of social network characteristics to provide a positive learning experience.

**Wikipedia:** It is a free encyclopedia that anyone can edit. Admittedly, Wikipedia is a quick and efficient way to search for history knowledge. However, since most of the content is generated by users, it doesn’t provide users with comprehensive and accurate history information. In addition, Wikipedia doesn’t support the capability of comparing different perspectives of one history fact and a collaborative learning environment to engage users, which is exactly what our system supports.

System Description

The core experience we envision for our users is to create collaborative timelines with friends, collect events as well as deploy discussions with them. Features such as compare lists, explore network, and jeopardy games, and so on, are designed to enhance both individual and collaborative learning experience. Users find historical events through browsing and searching. Since users’ learning activities are performed around timelines, any timeline creation or updates will be posted in the news feed.

Timeline Feature

The current way of learning history is not reflected in how well people manage their history knowledge and share it with others in certain ways. We borrowed the word “timeline,” reshaping it into a personal or collaborative documents managing tool. Users can create as many timelines as they like, area of where they would be able to collect historical events on a given topic, time period or simply an interest. Users can add any event they find on the site to their timeline. In addition, multiple users can edit one timeline. For instance, a user can create a timeline for World War II, and share with another two friends (Figure 1). By doing so, the user enables his two friends to edit the timeline with him simultaneously. They could add events about the war into the timeline, as they wish. In addition to this timeline, the user is allowed to create others; for example, a timeline named “movie camera” to collect an historical timeline of how movie camera have evolved.

On the homepage, to better serve this function, we provide fast access to the user’s timelines. By enabling them to do this, users will be able to drag the events showed in the news feed to their timelines. In addition to this way of creating a timeline all by oneself, a user can copy an existing timeline that their friends own.
Compare List Feature

The Compare list empowers users to explore the intersection areas between timelines. Users can set up two timelines concurrently; the purpose is to see interesting events that could have occurred at the same time. The timelines being added to the list could be the ones that users created, or those owned by others. Users can generate some thoughts about the comparison, and put them into the news feed. For example, a user might want to know the European economy in World War II. Using the compare list, they can simply add his timeline “European Economy” and the timeline “World War II” into his compare list and see what interesting results could be derived from the comparison (Figure 2). The way it works helps users to find out hidden connections between history events what other users put together. In addition, the compare lists can be saved if users wish to use them at a later date.

Explore Feature

For each keyword or topic users search, they can utilize the explore feature to find out more relevant content around the given topic. For instance, the event page is about WW II. When the user clicks explore, an infographic would be generated around this topic—WW II surrounded by concept bubble of Japan, China, General Eisenhower, Germany and so on (Figure 3). The connection and relevance between different concepts are captured in the distance between the concepts bubble and the thickness of the connection lines between the bubbles. Each bubble is clickable and may be placed as the center topic to refresh the infographic. Furthermore, in the concepts pool above the infographic, users can also refine their infographic by editing the concepts they wish to explore, either selecting from the recommended concepts or adding their own. By clicking each bubble, users could see a short description of the topic; while by clicking the edge between bubbles, they could see the explanation for the connections. Current system classified all the entities into three categories: people, locations, and events while the connections are briefly explained in text.
Gamification

As a way to evaluate the learning performance of users, our system provides a space where users could play Jeopardy-style games. The game offers two modes, one for individual players competing with each other and one for groups competing against other groups. The questions for group mode might be more complex than individual mode since they need collaboration within the group because each group member may only know one factor of the problem but they could come up with an answer working together. For example, a question might be asking about the world economy development in World War II. Someone might be only mastered of European economy, while someone knows the Asian economy. The best way to figure this problem out needs the collaboration of both of them. The group mode encourages users to find other users who are interested in the same topic, and learn from each other.
To meet the users’ needs of searching a specific concept, we provide a powerful searching functionality. First, the search results are listed by relevance. Next to that, we put an “Explore” button which could pop out a window that shows the connections between other concepts and this concept. More details about this function will be further discussed in the following section. Below that, there is an advanced searching tool provided to refine the search results (Figure 4). Users can check off any category they want to apply such as events, timelines, group activities, comments, people, etc. In addition, we enable search by location and time which are two key attributes in historical event exploration. Users can set the location either in the global map and or in the search box and set a time frame with up and down arrows in time axis and change the time interval. For example, as we selected Germany as our searching criteria, the searching results will be events about WWII that happened in Germany.

**Search Function**

**Evaluation**

We conducted a preliminary study of our system with five users who took part in our user testing on our hi-fi prototype. We randomly selected five students in University of Michigan Hatcher Graduate Library. To test whether our system is user friendly, encourages collaboratively history learning and offers an interesting learning process, we designed several tasks, including creating timelines, adding events to a timeline, comparing two timelines, and exploring an event related to a given event. We chose to focus on the qualitative performance of the subjects in the tests. All of the five participants completed these tasks successfully. In the post-test questionnaire, they reflected that compared to the current history learning resources (e.g. Wikipedia), the timeline feature did provide a better and more interesting way for them to manage the historical facts they like and collaboratively study history with friends. Jeopardy game, as one of the main social learning features, stood out as an attractive feature that users wanted to play around. They would like to play the game to compete with their friends. A higher score could make them feel proud. Therefore, our collaborative learning method was proved to be effective. However, more research is needed on the gamification because users could not actually play the game in the current
version prototype. In addition, two users indicated that the concepts of “timeline” and “event” were a little confusing because of the use of them to represent news feed in social network sites. In general, all the participants reflected that they would use our system and they liked the features.

![Advanced Search](image)

**Figure 4. Advanced search**

### Discussion and Future Work

Though our history learning site helps users collaboratively learn history effectively, it does have some limitations. The point mechanism is not a seamless way for evaluating the learning performance. Since points can be accumulated through posting comments, editing timeline and other social actions, it is possible that a user keeps leaving dummy or nonsense comments to increase points. Then the evaluation based on the points would be biased. In addition, our data have not been proved to be completely reliable and our database is not complete. Currently, all of our data of historical facts is gathered from Wikipedia.

According to the interviews conducted in the research stage, some interviewees reflected sometimes the information from Wikipedia is not sufficient and trustable. Also, our system evaluation is limited. We only interviewed and tested our main targeted users aged from 18 to 28. It is likely that other types of users like children and older adults hold different opinions from the interviewed users. In our future work, we need to enrich our database. We plan to cooperate with libraries and try to have librarians help us transform the historical records in the library to the electronic format, which then can be used as our data source. In this way, we can collect more reliable data so that users can find whatever they want to learn and trust them as well. In addition, we will gather more feedback from different kinds of users and improve our system based on their responses.
Conclusion

We introduce Timeline as a personal or collaborative documents managing tool, with which users can compare different timelines to explore the intersection areas and hidden connections. The multi-user editing will greatly influence the way people learn history. In addition, group activity and social games within group feature encourages users to discuss with others, share opinions, and merge ideas, etc. Furthermore, we develop the Explore feature, which generates an infographic to display the connection between different concepts, which enriches users’ personal learning experience to some extent. We also enable advanced search capabilities to help users to navigate a large collection of events, timelines, group activities, comments, people and more on our site, as well as recommendations based on the items. The social features at our site go beyond the sharing and commenting on the post, to allowing users to collaboratively create timelines and participate in a joint Jeopardy game.

References

UPitt iRiS Lab at iConference 2013 Social Media Expo: Search, Share and Learn the Way You Want

Zhen Yue
zhy18@pitt.edu
Wei Jeng
wej9@pitt.edu
Jiepu Jiang
Jij29@pitt.edu
Shuguang Han
shh69@pitt.edu
Daqing He
dah44@pitt.edu
University of Pittsburgh

Introduction

Social context and collaboration are among the important factors that motivate people to learn. Meanwhile, it is known that people learn best from interaction and collaboration with other people (Wlodkowski, 1999). As a matter of fact, people are often engaged in collaboration in workplaces or daily life not only because they need to handle the complexity of the tasks, but also because they can learn and socialize through the process. In information seeking and retrieval environments, when the search task is exploratory, it may be in the searchers’ best interests to collaboratively explore the information space and participate in shared learning (White & Roth, 2009). However, it is hard to have a one-size-fits-all system to support people engaged in various collaborative searches and shared learning due to the fact that collaboration can take many forms. Golovchinsky et al. (2008) summarized two types of collaboration among Web users. Implicit collaboration refers to a group of people who share similar search interests; the recommender system suggests items to these people based on the similarity of their interests. The goal is to use information previously found by others to inform new search results. Explicit collaboration, on the other hand, refers to a small group of people who share the same information need and who want to work as a team to engage in searching for information pertinent to the information need.

Figure 1: Collaboration in different social circles

People actually interact within different social circles when they engage in different types of collaborations. For instance, explicit collaboration often happens among people who are closely-connected such as family, friends or colleagues because they know each other personally. Implicit collaboration, in contrast, often happens among people who are aware that their results are based in part on data obtained from other users, but they may not know who the identity of the other users or what purpose the users had in mind while searching. We believe that factors such as search task or search context can motivate people to adopt different collaboration styles. Therefore, it is important to provide people with choices of system support so that they have the freedom to select the way they want to
search, share and learn information when interacting within different social circles. Those who are closely-connected may be comfortable using explicit communication and sharing of information, whereas loosely-connected people may garner more benefit by creating and maintaining community resources and collections. It is also important to enable people to reach out and explore potential collaborators. Our goal in this expo is to design a variety of components for our iRiSer system to support people’s freedom to search, share and learn in their preferred ways when interacting with people from different social circles.

**Searching, Sharing and Learning in a Team**

**Scenario**

Students may search as a group for information as part of a collaborative course project; friends may search together to find information to plan a vacation; healthcare providers may collaboratively search for information to diagnose a patient’s illness; or family members may collaboratively search on the Web to buy a car. All these scenarios share the same premise: that two or more people have the same information need, and they are working together in the same timeframe to satisfy that need. In other words, people search, share and learn information as a team.

![Figure 2: iRiSer System Team Component](image)

**The Team Component**

The team component in our iRiSer system is designed to support the aforementioned scenario. As shown in Figure 2, the iRiSer system’s interface has a left-hand panel dedicated to intra-team communication – the Chat Box, and the main panel is further divided into three frames: topic statement, Web search and team workspace. The topic statement frame contains the description of the task on which the user is currently working. The search frame connects the user’s query to Google and displays the Google search results. Users examine search results for relevant information, and can save a whole Web page or a snippet of the page. Users can also see their search history (queries) as well as those of their teammates. The team workspace is designed to facilitate the sharing of relevant search results. Users can click to view more details about an item in the workspace, as well as to comment on or assign tags to any item. Users can also decide whether or not certain items will be visible to other team members. Overall, this component provides the team members with the opportunity for coordinated searching: they have the same topic statement space, they can use IM to coordinate their search process, and they can share search products or learn from each other’s activities.
Searching, Sharing and Learning in a Community

Scenario

People in the same community often share similar information needs and interests. For example, residents of the same city may be interested in learning about the local news and events in that city. People in interest-oriented communities, such as those who play the Starcraft computer game, may be attracted to the same new updates to the game. People from the same profession are all interested in the latest breakthroughs and emerging topics in the discipline. However, as individuals in the community, each person can also bring different perspectives to the otherwise homogeneous landscape. This helps to establish the variety inside the community, thus making it interesting and important to characterize the similarities and differences among different people. This provides heterogeneous and rich contextual information for searching, sharing and learning within a community.

The Community Component

The community component in our iRiSer system supports the community scenario (See Figure 3). In this component, people’s search queries are used to not only match relevant scientific articles, but also to locate online users in social reference management websites (such as Mendeley and CiteULike) with similar interests as well as the authors who write and cite articles in similar topics. Thus, we create a virtual community that contains the users and authors who share the same interests as the searcher. The search results are optimized to consider multi-dimensional relevance criteria including: topical relevance scores, as calculated by matching the queries with article content; community reading interest scores, which measures the extent to which the article can satisfy the interests of this virtual community; and academic publication contribution scores, which measures the likelihood that scientists, when writing articles on similar topics, would cite the article. Considering all of these factors, the system models scientists’ information needs by incorporating the contexts of communities in which scientists are interested.

Discovering Potential Collaborator

Scenario

Despite having access to numerous web documents returned by powerful search engines, users on many occasions still feel that it is more important to be connected to the right people (Ackerman, Pipek, & Wulf, 2003). There are various reasons that people seek out new collaborative relationships. For instance, in the academic environment, researchers may want to find a collaborator to work on a cross-
disciplinary project, doctoral candidates may need to find external committee members, and conference program chairs may need to look for committee members. Each of these tasks and contexts may have significant influence on the collaborator discovery process.

The Collaborator Component

The collaborator component we built into our iRiSer system supports the finding of appropriate collaborators. Most searches for people are exploratory in nature, in which the users may only have vague ideas about whom they are seeking. Therefore, it is optimal to model people discovery as an interactive exploratory process, in which the users and the system are engaged in an iterative process of stating the requirements and exploring the returned candidates. Figure 4 reveals the interactive exploration process of finding collaborators for the topic of “digital library”. The default factor is content-based relevance, but the system also incorporates two other factors: authority and social similarity. Authority measures the author’s reputation, which is calculated using PageRank algorithm. Social similarity measures the similarity between candidate and the searcher in a social network. Returned candidates are presented with their names, affiliations, and the scores for each factor. Users can issue queries to define the search topic, adjust the importance of each factor (by sliding its bar to generate updated candidate rankings), and save the relevant candidates in the workspace shown in the right hand panel.

Future Work: Integration

Each of the three iRiSer system components described in this paper were developed separately and evaluated in several user studies. There were several critical lessons learned during the studies: our system can support users in different social contexts for exploring a wide range of information spaces, and users are happier and more satisfied because they can socially engage with others rather than working alone. However, we are aware that users’ needs are often multi-faceted and dynamic in nature. Unfortunately, each system component in isolation can only support one such facet or take a static view of people’s social circles and collaboration relationships. Therefore, we are in the process of integrating these components into one seamless service to provide users with the freedom to choose the way they want to search, share and learn information. The two challenges that we face are 1) how to make smooth transitions among different components; and 2) how to automatically discover users’ needs and provide the appropriate support?
References


The Need for Information Gurus in Online Learning Communities

Nick Josten
iSchool@UBC
nickjosten@gmail.com

Jonathan Kift
iSchool@UBC
daftkifty@gmail.com

Allison Trumble
iSchool@UBC
allisonptrumble@gmail.com

Brigid Winter
iSchool@UBC
winterkist@gmail.com

Abstract

This study argues the need for So.cl to identify and promote information gurus within their online learning communities to aid informal learning and foster communication between communities.

Keywords: affinity spaces, community of practices, expertise, information brokers, privacy

Introduction

Informal learning communities are comprised of individuals passionate about certain topics of inquiry. This passion causes people from various backgrounds, geographical regions, and areas of expertise to gather and exchange ideas and encourage each other to grow and learn. Within these communities relationships form, experts mentor non-experts, and members interact with outside communities. Microsoft’s So.cl system provides a platform for such communities to develop and share information. However, the system emphasizes personal search and inquiry, and does not fully tap into this passionate communal dynamic. After an examination of relevant literature as well as the dataset provided by Fuse Labs, combined with our own explorations of So.cl, we propose several additions to So.cl that increase the findability and accessibility of what we have coined “information gurus”. These are experienced members who already exist within So.cl, who are contributing valuable and informative content and can liaise between other members or communities. In this paper we present the results of our research and analyses in support of our suggestions for So.cl.

Passionate Affinity Spaces & Communities of Practice

Passionate affinity spaces arise when individuals share particular interests and come together around them (Gee & Hayes, 2011). These individuals first find each other as a result of their shared interest, but continue to interact and build relationships because of a common learning goal. While they are first brought together through the topic, their participation in a learning community develops into a sense of responsibility towards that community, and in turn shapes the way they contribute to it (Riel & Polin, 2004). This sense of responsibility or accountability is what continues to drive the community to engage in learning and in sharing that information with one another. With “a strong emphasis on the notion of a community as a shared activity and goals” driving the community to learn, explore, and explain their interest to one another, the identities of the members begin to be defined by their very role in the community (Riel & Polin, 2004). Since individuals are already passionate about their shared interest, they have more of a stake in how the community develops, as “the things we say, the things we choose to read or view, the things we link to, the people we send messages to - all of these constitute input to the learning network, causing it to reform” (Downes, 2010). While individuals may search for information and learn based on their own self interest, involvement in a community encourages them to pursue the topic further while providing outlets for expression and guidance to new information. Indeed, it is through their participation in the community– through sharing information and interacting with others, that real learning actually occurs, rather than simply through the ingestion of knowledge (Riel & Polin 2004).

Acknowledgements: Fuse Labs and iSchool@UBC for funding, thank you to Elizabeth Schaffer, Luanne Freund, and Rick Kopak for your support and guidance, and to LIBR 514F: Visual Analytics.


Copyright is held by the authors.
These affinity spaces and communities pop up online, allowing self-motivated and self-driven individuals to engage in learning. As Riel & Polin (2004) describe there are quite naturally some members who have more expertise than others, but how can newcomers identify them? As So.cl continues to grow as a social space supporting online learning, what will enable new users to delve into their communities of interest and find other communities discussing related topics they might find interesting?

**Information Brokers or “Gurus”**

Through interactions within communities, strong and weak ties develop (Haythornthwaite, 2002). Strong ties lead to more communication within a community, but the information being shared may be redundant, as similar information is shared and re-shared. Weak ties with people interacting with other communities provide access to new information. From these interactions, information brokers arise (Haythornthwaite, 2002; Wenger, 2000). These brokers move between the community boundaries to forge connections with “groups engaged in tangentially related practices” (Riel & Polin, 2004), which keeps the information fresh and prevents enclosed communities from recycling the same information and practices.

If users can find these brokers, or “information gurus”, they can more easily find the connections between topics of interest and pursue relevant information that may have gone undiscovered. Since the information gurus naturally arise within communities, So.cl can use metrics to identify these users and promote them within the platform, allowing inexperienced users to move between communities and find more relevant information. Possible ways of indicating an information guru include tracking comments and riffs on posts to measure the impact the potential guru has within a learning community. Other possible metrics are discussed further in the “Gleanings from the Dataset” section.

**Expertise and Mentorships in Communities**

While acting as a liaison between communities, information gurus have a secondary benefit of acting as a mentor within communities. If So.cl develops a way to acknowledge experts and make them easily findable, users can locate more quality and relevant information within the community. The benefit of affinity spaces and communities of practice is that they function through an apprenticeship model, in contrast to a formal teacher-student structure: “social learning involves acquiring the practices and behavioural norms of established practitioners within a community of practice, a skill which has historically been acquired through an apprenticeship or a mentor-mentee relationship” (Grewal & Harris, 2009). All members of a community contribute to the group – either actively or passively – and in so doing, create the community's knowledge base. However, new members often require guidance, which usually comes from experienced members, as in communities of practice that rely “on tacit understandings that are shared among members and passed along through mentoring and apprentice experiences” (Riel & Polin, 2004).

**Gleanings from the Dataset**

The dataset of So.cl interactions contains information about the topics someone posts about, the people who follow them, and the interests those people follow, among other things. We can choose a common interest, such as art, and look at who is following that interest, who is creating posts tagged with that interest, and whether the people following the interest are also following the individual posters.

A visual network analysis of the so.cl dataset was conducted as part of this project. The dataset was quite large and needed to be limited in some way in order to derive meaning from it using a visual analysis. Though the dataset included records for a full calendar year, only records between June and November of 2012 were included, as these were created after Microsoft lifted the invitation-only restriction on membership. As one of the objectives was to examine how users interacted around a particular topic, it was necessary to choose a topic to focus on. The topic “art” was chosen, as it is one of the first suggested interests when users first join so.cl and it contained enough records to generate some insight.
The analysis required the dataset to be significantly restructured in order to select the desired records and generate calculated fields necessary for the metrics to be examined. Both MS Access and Excel were used for this manipulation. NodeXL was used to create the visualizations used in the analysis.

The following visualizations depict the relationship between users who create posts tagged “art”, and the users who follow them. All of these users also follow “art” as an interest. The first visualization pictured here shows this subset with node colour representing indegree, which in this case represents followers, and node size representing the total post count associated with that topic.

Fig. 1: Users following people posting about art. Colour represents indegree or total followers and node size represents total posts by each user. Both were calculated logarithmically and had outliers ignored.

This graph shows that most of the smallest dots are very orange, indicating that those who do not post much do not attract many followers, while most of the biggest dots are very blue, confirming that those users who generate the most posts also get followed the most. This indicates a very strong correlation between indegree and post count, which suggests that users are generally successful at finding the users who are posting the most about the topic they’re interested in. But are these the most valuable posts? One way to measure whether the posts were actually of any interest to anyone was to calculate the average number of comments per post each user receives. This would at least indicate that the post is driving some sort of engagement. The second graph uses this calculation as the size of the node, with indegree still represented by colour.

Fig. 2: Users following people posting about art. Colour represents indegree or total followers and node size represents average number of comments per post for each user. Both were calculated logarithmically and had outliers ignored.

Whereas the previous graph showed an almost perfect correlation between colour and size, this graph shows several large nodes that are quite orange, meaning that there are several users who generate a high number of comments per post but did not have very many followers. If comments per post can be relied on as a measure of quality, then this means there are people who are contributing quality information to the learning community but people interested in that topic are not finding them. This is a problem that the designers of so.cl should address if they want to foster online learning, as currently users only seem to be finding others who are posting quantity rather than quality.
In all likelihood, however, comments per post is not a very reliable measure of quality. There is currently no clear way of ascertaining who is actually adding value to the conversation about art, or who could potentially be designated a “guru”. A guru metric could be developed by performing further calculations to combine comments per post with other measures currently collected by the system, such as the number of distinct users who comment on a guru’s post, or the number of the guru’s followers who also follow the interest topic they’re posting. This metric would be made much richer, however, with the addition of a simple way of “upvoting” posts, like a “This taught me something” button. More than a simple “like” button, this would be a way of indicating that the post actually contained some valuable or informative content that contributed to learning.

Even though much of this information is currently available to researchers, it is not available to the users. The incorporation of meaningful suggestions for people to follow based on the interests they are following or posting about could potentially facilitate the development of learning communities by connecting people to the “gurus” of those communities.

**Conclusion**

Information gurus can act as both mentors within communities and brokers between communities. While experts and information gurus already exist within So.cl, they are not adequately showcased. By allowing users to vote on the quality of posts and combining this into a metric that includes comments per post and number of followers, So.cl can identify those users who can most effectively link communities and provide expertise and mentorship within communities. This metric could recognize who is contributing valuable information to an interest community and reward information gurus while helping others find them. While the analysis done with this dataset shows that users can more effectively link to one another through gurus, further study could be conducted into the possibility for automatically created tags or suggested tags and retroactive linking between riffs and posts, which could add a richer connection between users and gurus. This could potentially enable gurus to further activate their roles as information brokers between online learning communities.

**References**


