The Association for Computers and the Humanities
The Association for Literary and Linguistic Computing

Digital Humanities 2007

The 19th Joint International Conference of the Association for Computers and the Humanities, and the Association for Literary and Linguistic Computing
University of Illinois, Urbana-Champaign
June 4 - June 8, 2007

Conference Abstracts

Graduate School of Library and Information Science
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Graduate School of Library and Information Science

International Program Committee Members

- Jean Anderson (ALLC)
- Elisabeth Burr (ALLC)
- Espen Ore (ALLC)
- Kevin Hawkins (ACH)
- David Hoover (ACH)
- Ray Siemens (ACH; Chair)
- Paul Spence (ALLC; Vice-Chair)
- Christian Wittern (ACH)
- Natasha Smith (ACH)

Local Organizers

- John Unsworth, GSLIS
- James Onderdonk, Jr., Conferences and Institutes
- Susan Etter, Conferences and Institutes

Editorial Team

- Sara Schmidt
- Ray Siemens
- Amit Kumar
- John Unsworth

ISBN: 0-87845-125-0
Second Edition, incorporating corrections to the original printed edition distributed at the conference
Published by
Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign
Conference logo by Larry Steinbauer
Set in Liberation Serif, designed by Ascender and distributed by Redhat under GPL

Introduction

Welcome to The University of Illinois at Urbana-Champaign—the birthplace of ILLIAC (the first computer built and owned entirely by an educational institution), PLATO (the first online instructional program and home of the first online community), The Illiac Suite (the first piece of music produced with a computer), Project Gutenberg (the first online collection of literary texts), Mosaic (the first graphical web browser), and Richard Powers (author of *Galatea 2.2* and *The Gold Bug Variations*). UIUC is also home to the National Center for Supercomputing Applications where computer engineers and domain experts pursue grid computing, security, visualization, datamining, scientific computing, and petascale computing. NCSA is a sponsor of this conference, along with CHASS, the new and NCSA-based Center for Computing in Humanities, Arts, and Social Sciences. The other sponsor, and home to the local organizer, is the Graduate School of Library and Information Science, or GSLIS. GSLIS (pronounced "gislis") has the oldest doctoral program of its kind in the U.S., and it offers a professional masters program that is ranked in first place nationally, as well as an advanced graduate degree for practitioners in digital libraries and an undergraduate minor in information technology studies. GSLIS has been delivering online degree programs since 1996, and it currently enrolls half of its students in the online program. The GSLIS faculty includes some well-known names in humanities computing, including Allen Renear, past president of The Association for Computers and the Humanities.

Many people contributed to the materials that are represented in this book of abstracts. As local organizer, I would like to begin by thanking two people whose diligence and effort made this conference possible: Ray Siemens, who chaired the international program committee with great aplomb and efficiency, all the while making it look easy, and Sara Schmidt, who assisted in every aspect of the local process, from registration to submissions for review, to exhorting authors to turn in final copy of accepted abstracts, to marking up the abstracts into their final form for this book and their presentation on the Web. Sara has thrown herself into organizing this conference with a will, and she has found a way to make everything work.

So many others need to be thanked, beginning with the authors who submitted proposals to the conference in the first place, and to all of those who contributed their time and expertise to reviewing submissions to the conference: the quality of what we present at this year's conference is due to both kinds of effort. Thanks also go to the International Program Committee, who—under Ray's able leadership—vetted the outcome of the reviewing process and made the difficult final decisions on the program. I'd also like to thank Harald Weinrich, of Hamburg, Germany, developer of the ConfTool system that we used this year to manage the reviewing and registration processes: he was very helpful when we needed changes to that system, and I recommend him and ConfTool highly.

Thanks also must go to Martin Holmes, for helping us to re-use the XML/XSL-based abstract publishing system developed for the 2005 iteration of this conference, in Victoria, British Columbia, and Amit Kumar for getting that system set up under digitalhumanities.org, where it can be re-used in future years, and for tweaking the stylesheets and other aspects of that system for this year's conference. David Dubin and Kevin S. Hawkins helped with translation issues, from German and Russian, respectively, and Christian Wittern, Elisabeth Burr, Peter Liddel, Alex Bia, and others helped in extending and adapting translations in the ConfTool interface; Syd Bauman and Julia Flanders helped to answer questions about TEI encoding of the abstracts; and Hana S. Field assisted Sara in marking up those abstracts. Melissa Terras helped in rounding up information about current membership/subscribership, from Oxford University Press, and Miranda Remnek helped with bursaries for our Eastern European colleagues. Several of those were very generously paid for by Vernon Burton, out of the resources of the Center for Computing in the Humanities, Arts, and Social Sciences here at the University of Illinois, Urbana-Champaign. Some were also covered by the Associations (ACH and ALLC): this is one of the many worthwhile things that the Associations do with the income that individual subscribers provide.

I hope you enjoy your visit to Champaign-Urbana, and I hope you have a chance to take the excursions to Springfield (the Lincoln Museum and the Dana Thomas House, a Frank Lloyd Wright commission) and Allerton Park (recently declared one of the 'seven wonders of Illinois'). In between, I hope you have a stimulating and successful conference, hear many fine papers, and engage in those hallway conversations that are sometimes the most rewarding part of a small conference.

John Unsworth
Local Organizer

About this abstract collection

Planning for Digital Humanities 2007 began in earnest in the Sorbonne's Maison de la Recherche, in July 2006, at the first conjoint conference of the Association for Computers and the Humanities (ACH) and the Association for Literary and Linguistic Computing (ALLC) under the umbrella of the Association for Digital Humanities Organisations (ADHO) – a meeting most certainly inspired by the collegiality, congeniality, and grace of our hosts, Liliane Gallet-Blanchard and Marie-Madeleine Martinet of CATI, the research centre for Cultures Anglophones et Technologies de l'Information. With the success of our time together in Paris as exemplar—the Parisian gathering itself modeled on successes of the past at Victoria, Göteborg, Athens, Tübingen, New York, Glasgow, Charlottesville, Debrecen, Kingston, Bergen, Santa Barbara, Washington, Tempe, Toronto, among many other notable loci for our community—near the end of a conference that we roundly acknowledged as being a high water mark for our community, we had good reason to anticipate with considerable optimism our next community event: the 2007 meeting in the 'Paris of the midwest', at the University of Illinois Urbana-Champaign.

As this brief rehearsing of important gathering places, people key to our recent events, and the research centres kindly sponsoring our conference activity rightly begins to suggest, no conference organizing group works in isolation, nor without a helpful context. Indeed, in a community like ours, conference organization can be as much a pleasure as conference participation, for each is an activity involving the interaction of members of the community—from the conception of the annual conference theme (which often arises out of key points of focus from earlier conferences), to the breaking-in of a new conference management tool for our community (which involved absolutely everyone), to receiving and reviewing presentation proposals (involving a group numbering circa one hundred), to the conference itself (which has an impact beyond the conference attendees, eventually on thousands of interested individuals).

While we all participate, some notes of special thanks are due to our local hosts and sponsors, represented most generously in the person of John Unsworth, and to those involved in the conference assistance from which we all benefit, personified with utmost efficiency by this volume's co-editor, Sara Schmidt. A note of especial gratitude is due to all those who so diligently participated in the review of submissions for the conference: Adrian Miles, Allen H. Renear, Amy Bruckman, Anna Bentkowska-Kafel, Bethany Nowviskie, Bruce Robertson, Christian Kay, Christian Wittern, Christie Carson, Christine Ruotolo, Chuck Bush, Claire Warwick, Claus Huitfeldt, Dan Tufis, David Bearman, David G. Durand, David Gants, David Green, David L. Hoover, David S. Dubin, Doug Reside, Elisabeth Burr, Elli Mylonas, Espen S. Ore, Federico Meschini, Francisco Javier Carreras Riudavets, Gabriel Egan, Gary F. Simons, Gary Shawver, Hanno Biber, Harald Baayen, Hazel Gardiner, Hugh Craig, J. Stephen Downie, Jay David Bolter, Jean Anderson, John

Bradley, John Dawson, John Lavagnino, Joseph DiNunzio, Joseph Rudman, Julia Flanders, Karen Wikander, Kim Plofker, Lars Johnsen, László Hunyadi, Lisa Hopkins, Lisa Lena Opas-Hänninen, Lou Burnard, Lyman Gurney, Manfred Thaller, Marie-Maddeleine Martinet, Mark Olsen, Marshall Soules, Martha Nell Smith, Martin Holmes, Martyn Jessop, Matthew Kirschenbaum, Matthew Zimmerman, Melissa Terras, Michael Neuman, Michael Sperberg-McQueen, Nancy Ide, Natalia (Natasha) Smith, Neil R. Fraistat, Øyvind Eide, Patrick Conner, Patrick Juola, Paul Caton, Paul Joseph Spence, Penelope Gurney, Perry Willett, Ray Siemens, Richard Gartner, Richard Giordano, Ron Van den Branden, Sebastian Rahtz, Stéfan Sinclair, Susan Brown, Susan Hockey, Susan Schreibman, Syd Bauman, Tanya Clement, Wendell Piez, Willard McCarty, William Kretzschmar, William Winder, Worthy N. Martin, and Zenón Hernández Figueroa.

With very best wishes for a pleasant and fruitful time together!

Ray Siemens
Chair, International Program Committee

Table of Contents

Index of Abstracts

Abstracts

QRedit: An Integrated Editor System to Support Online Volunteer Translators

Takeshi Abekawa *(abekawa@p.u-tokyo.ac.jp)*
University of Tokyo
Kyo Kageura *(kyo@p.u.-tokyo.ac.jp)*
University of Tokyo

1. Introduction

We are currently developing an English-to-Japanese computer-aided translation (CAT) system with the aim of aiding online volunteer translators, who are involved in translating online electronic documents in their free time. In the following, we will first describe the characteristics of target users, and discuss the basic concept of the system within which the choice of functions and system specifications has been made. Then we will describe the prototype CAT system which we are now experimentally providing to a limited number of online translators.

2. Background and basic concepts

Many CAT systems have been developed to date [see Kay 1997 for background information and Gow 2003 for a comparison of several CAT systems]. While some existing CAT systems have proved useful for some translators and translation companies, volunteer translators who work online do not use these systems [Kageura, et. al. 2006; see also Fulford and Granell Zafra 2004 in relation to the situation of freelance translators in Britain], for a variety of reasons: they are too expensive for personal use (except for Omega-T, which is free [Omega-T 2007]); they have more functions than are necessary; such translators are not under pressure from managers to keep a translation log in order to control quality, etc. This shows that some sort of CAT-lite system is required by online translators, which shares some basic functionalities with existing commercial CAT systems but with a different emphasis and overall design principle [Boitet, et. al. 2005].

We have therefore been engaged in developing a CAT system for online volunteer translators, based on the principle of maximally aiding the translators' work flow by removing existing obstacles, rather than providing extra functions which translators have never used. Thus we are taking a bottom-up approach in developing system functions, reflecting the concrete requirements of translators. Another essential principle we have adopted is that it should be translators who make decisions, not the system.

After consulting some 15 volunteer translators working online, we identified the following issues as being particularly important:

1. Most English-to-Japanese volunteer translators do not have native level command in the source language (English). As a result, unlike such CAT systems as TransType [Macklovitch 2006] which assume that the target users are bilingual, well-trained and professional translators, and therefore try to reduce the time used for inputting the target language text, the main point to be tackled in our environment is making the reference lookup process as easy as possible.
2. When selecting translation expressions, translators examine various possibilities. When necessary, they look up more than one dictionary and check other information resources such as print encyclopaedias or Wikipedia, etc. The CAT system should facilitate the translation process by reducing the effort involved in looking up multiple resources. As it is not the mission of the system to decide on behalf of translators, good output for the system is considered to be a range of candidates and information that translators can take into account in making decisions, and not simply the same expression as translators chose.
3. Some online translators do not use online dictionaries, because doing so breaks the rhythm of translation and hinders the construction of the target text. Partly because of this, some prefer using print or standalone electronic dictionaries. A few use dictionaries built into the text editor, but none of them are fully satisfied with the dictionary look-up environment.
4. There is a pressing need for improved idiom and phrasal look-up. The importance of this function derives from two factors: (a) many translators, even experienced ones, have relatively less knowledge of idioms than of words, and (b) some idioms may not be identified as such by translators, because they make sense without an idiomatic interpretation. This leads to translation mistakes.
5. As our system does not aim to choose and restrict information on behalf of translators by using "sophisticated" NLP techniques, the amount of information it provides will naturally tend to increase. As a result, the user-interface becomes an important issue.

3. The prototype system QRedit

Based on these concepts and requirements, we are currently developing a prototype system that supports online translators. The system runs on the TOMCAT server and users access the system through a Web browser. The overall image of QRedit is shown in the figure. The browser screen is divided into two areas: (i) the source text area, and (ii) the target text area. The users can choose between horizontal and perpendicular division of the areas, i.e. source text area on the bottom and target language area on the top, or source text area on the left and target language area on the right. The two areas are linked with a synchronised scroll function.

3.1 Functions in the source text area

After a translator inputs the URL of a Web page (in which case the system analyses the tags and extracts the textual area automatically) or copies and pastes text into the source text area, the system activates the dictionary look-up functions. When the user clicks on a particular word in the source text area with the mouse, the system shows the translation candidates in a pop-up window.

Dictionary look-up functions

The system displays translation candidates from the dictionaries incorporated into the system [Sanseido 2004; Eijiro 2006]. The system does not only incorporates simple word look-up functions but also incorporates flexible idiom look-up functions. The idiom look-up functions can match such idiom occurrences as "He said that with his big fat tongue in his big fat cheek" with the dictionary entry "with one's tongue in one's cheek." This function has not been realised in any English-Japanese MT systems we have checked, and while some CAT systems realise similar functions through approximate matching, they do not specifically target the look-up of idioms with their variations. The system alerts users to idioms by marking them with an underline.

Displaying translation candidates

The system can display information in two ways. One is a small pop-up window displayed within the source text area, which includes only target word candidates. The other is a large pop-up window that displays all the information given under the headings in the dictionaries. The latter is particularly useful when translators wish to examine related information in detail before deciding on a translation, while the former is convenient for more straightforward dictionary look-up.

3.2 The connection between the source text area and the target text area

Translators can paste a selected expression from the list of candidates in the pop-up area to the target text area by clicking the mouse. The system also provides automatic transformation of numerical expressions to Japanese conventions and the pasting of HTML tags. This mouse operation does not affect the keyboard operation, and whatever the user does with the mouse, the keyboard cursor always stays on the target text editing area, so that the user can input translated text continuously.

3.3 Target text area

The target text area consists of the open source Web editor FCKeditor, which offers WYSIWYG textual decoration and editing, as well as saving and loading functions. The text can be saved in HTML as well as in basic textual format. The target area is split into paragraph spaces which correspond to the source text.

4. Conclusions

We are currently running a prototype system of QRedit that incorporates these functions, and accumulating feedback from a few online volunteer translators. The feedback that we have obtained so far can be categorised into two types:

1. The need to refine existing functions and improve basic usability, e.g., to improve the accuracy of the extraction of the textual area when the user specifies an URL for the source text area.
2. The need to incorporate higher-level functions, e.g. to allow the user to specify the register of the text that (s)he is translating, in order to have the system block irrelevant dictionary information from being displayed.
3. In addition, we are developing a module that automatically compiles bilingual technical terminologies from the Web, a module that detects existing translation pairs and recycles the translation information, and a system that detects candidates for transliterated expressions of proper names from the Web, all of which will be integrated into the QRedit environment.

Bibliography

Boitet, Christian, Youcef Bey, and Kyo Kageura. "Main Research Issues in Building Web Services for Mutualized,

Non-commercial Translation ." *Proceedings of the SNLP: 6th Symposium on Natural Language Processing* . 2005.

Eijiro. Accessed 2006-10-31. <http://www.eijiro.jp>

FCKeditor. Accessed 2006-10-31. <http://www.fckeditor.net>

Fulford, Heather, and Joaquin Granell Zafra. "The Uptake of Online Tools and Web-based Language Resources by Freelance Translators: Implications for Translator Training, Professional Development, and Research ." *Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training* . 2004. 37-44.

Kageura, K. et. al. "Improving the Usability of Language Reference Tools for Translators ." *Proceedings of the 10th of Annual Meeting of the Japan Association for Natural Language Processing* . 2006. 707-710.

Kay, Martin. "The Proper Place of Men and Machines in Language Translation." *Machine Translation* 12.1-2 (1997): 3-23.

Macklovitch, Elliott. *TransType2: The Last Word*. LREC, 2006. 167-172.

Omega-T. 2007. <http://www.omegat.org/omegat/omegat_en/omegat.html>

Sanseido. *Grand Concise English Japanese Dictionary*. Tokyo: Sanseido, 2006.

Citation Networks: A New Humanities Tool?

Almila Akdag *(alelma@ucla.edu)*

Department of Art History
UCLA

Zoe Borovsky *(zoe@humnet.ucla.edu)*

UDHIG
UCLA

Abstract

This paper proposes to use citation networks as a tool for mapping out links among disciplines and research areas in & around humanities. By collecting papers from main journals in diverse areas like cognitive science, art history and psychoanalysis, this study aims to build a small citation network that will be visualized and published as a 3d web page. The end-map will be the backbone of a dissertation, which tries to analyze the influence of various movements and disciplines onto the art historical canon. Used as such, a citation network can become more than a frozen map that has more than a one-time use value.

Access to online electronic databases and tools such as Google scholar have led to a significant improvement in the discovery of secondary literature. However, organizing the vast amount of bibliographic information from various discipline-specific databases continues to be an impediment to truly interdisciplinary work. My project attempts to construct virtual maps of electronic databases or digital libraries capable of providing scholars with significant links between disciplinary relations, interdisciplinary research areas, or tendencies, approaches and methodologies inside a single discipline. Moreover, as I will show, if the third dimension, namely history, is added to these maps, the transformation of the disciplines, the merging of research areas, and the changes in the taxonomic structure of the academy will be revealed to the expert eye. As an example of how such a virtual time-map could become a valuable analytic tool during the research process, I will construct such a tool and, using my dissertation as a test case, demonstrate how it can be applied to map intersections between art historiography, psychoanalysis and cognitive science.

As a PhD candidate in UCLA's Art History program, my dissertation traces the changes of the critical discourse from

the 1970s, when the so-called "New Art History" clashed with traditional art historians, and gave rise to a whole new approach—one that has now become known as Visual Cultural Studies. However, in such a broad context, one needs to handle a deluge of texts and interrelations. A simple timeline, a linear outline constituted of chapters and subchapters is not enough to depict the map of overlapping layers, concepts and relations between opposite but—in this case still—allied methodologies. In order to render all these visible, I would like to create a multidimensional space of such relations. To that end I propose to collect significant papers, extract citation information using various text-analysis programs and visualize the end results as a citation network that runs along three different trajectories, namely the history of psychoanalytical and cognitive scientific methodologies and their impact on the evolution of New Art History into Visual Cultural Studies.

History:

A relatively new research venue, called 'scientometrics' or 'bibliometrics', specializes in creating such maps for delineating growth, relations and interactions in scientific fields.[1] Bibliometrics uses text analysis to extract citation data from papers and makes use of this data as a way of evaluating scientific publications. To obtain information from scientific papers and using he results in mapping out scientific relations has a long history. As early as 1964 Garfield and his colleagues suggested using citation data to evaluate the development of science.[2] From 80's on, the research in this area accelerated with the advancement of computers and various combinations of statistical methods used to extract and evaluate information such as citations, co- citations with reference of various bibliometric data. The end-results are usually rendered as so-called 'citation networks' which are a variation of social networks.[3] Now it is a common practice to evaluate a scholar or a journal according to how many times it/he/she is cited. Moreover, there is so much research done using citation networks as a methodology to analyze the disciplines, many scholars are now questioning the efficacy of this approach.[4]

Among the ample publications in this area, two general approaches distinguish themselves: the citation networks are either built to support an idea or to enhance the way in which such networks are composed. In the first instance, a search is done to filter out unnecessary papers. The maps generated in this way are limited with a scope, and generally give an overview of the research topic. These maps are not created from a relational database and are not published for further use. That means they are not applicable to other research questions and the enormous work put into collecting papers and preparing them for network analysis is done on a case-by-case basis because access to the databases is restricted and therefore the datasets themselves cannot be made available. In the second instance papers are extracted from the electronic database without any filtering. Instead a time limit is imposed. These types of publications tend to focus more on the technical details and explore the mathematical substructure of social networks. Usually open-source databases are preferred, since the main idea is to test the application's performance on huge datasets. Papers about such studies do not interpret the resulting map and instead detail the mathematical innovation of the applications.

Despite the popularity of this approach in the sciences, I have yet to find a paper that uses Humanities databases. Even the most comprehensive citation network, a data set that encompasses "7,121 journals covering over 1 million documents in the combined Science Citation and Social Science Citation Indexes" does not delve into the Arts and Humanities Citation Index.[5] Rather than simply applying the same techniques to Humanities materials, a fresh approach, one that is not only suits to humanities scholarship, but addresses some of the issues raised in the scientific communities seems desirable.[6] Although my long-term goal is to create a dynamic citation network that can become a part of an ongoing research project in digital humanities, in this paper I will focus on creating a static citation network using open-source tools. My research plan to achieve this aim is as follows:

1. *Collecting papers:* I have already collected around 2000 papers, mostly from prominent journals in Art History (such as Art Bulletin, October, Art Journal and Leonardo), and in Cognitive Science (Trends in Cognitive Sciences). Beside these resources I would like to include the classical texts in psychoanalysis; which can be found in electronic format in the Psychoanalytic Electronic Publishing database. The main criterion used for selecting pertinent papers is to use keywords that are relevant for the research topic. For example I used keywords such as "Freud", "psychoanalysis", "aesthetic", "artist" while searching in a cognitive science journal whereas I chose keywords like 'cognition", "cognitive science", "vision", "artist" etc. for the search in the domain of psychoanalysis.
2. *Preprocessing the database for text analysis & extraction the needed information:* The acquired corpus is preprocessed for text mining and analysis. Preparing a list of keywords and bibliometric data (author name, date, journal name, title, etc.) will be enough for the preprocessing stage. We will experiment with different text- mining and text analysis programs and report on those that work well at this stage of data preparation.
3. *Construction of social networks:* A social network is a graph representation of social relations. Graphs are the most popular and widely researched data structure for representing and processing relational data. In a graph, each node represents one entity (a person in a social network; a researcher or a work in a citation network) and the edges

(or arcs, if they are directed) of the graph represent some relation. One can also indicate the strength of the relation by associating weights with the edges of the graph. Then, by using tools like Pajek[7], the graph nodes can be placed in a 3D space in such a way to minimize an energy term. Thus, the nodes that are close to each other semantically (through the interpretation of graph edges) are placed in proximity, even when they don't have actual links. On a social network, clusters and cliques can be identified, indirect relations can be uncovered, and relevance judgments can be made based on quantitative or qualitative measures. Even the location of the nodes (center vs. periphery) can be informative for a person thoroughly acquainted with the represented structure.

The use of such a graph tool is simple; nodes and arcs are read from a file, and the graph visualization is accomplished with a few commands. For a citation or semantic network, text mining tools can be employed to derive the entities in relation automatically. Once such a network is built, Pajek can import the graph in 3D file formats like 3xd and VRML; both are now becoming standards for internet publication in 3D.[8]

Conclusion

The proposed tool serves not only as a bibliographic aid, but will become the main framework for my dissertation. By incorporating the 3D virtual map into my dissertation, I hope to demonstrate a new form of digital scholarship--one that springs from the new possibilities that digital technologies affords scholars in the humanities whose work is inherently, and exuberantly, interdisciplinary.

---

1. To read more on the history of scientometrics see Katy Börner, Jeegar T. Marus, and Robert L. Goldstone. "The Simultaneous Evolution Of Author And Paper Networks," *PNAS* 101 (suppl.1) (2004): 5266-5273.
2. All these search engines providing citation index information are products of Thomson ISI. The original foundation was called simply Institute for Scientific Information. Garfield launched it, again in 1964, see ibid.
3. Doreian Patrick, "A Measure Of Standing Of Journals In Stratified Networks," *Journal of the American Society for lnformation Science* 8.5-6 (1985): 341-363.
4. See R. E. Rice et al., "Journal-To-Journal Citation Data," *Scientometrics* Vol. 15.3-4 (1989): 257-282; Lindsey D, "Using Citation Counts As A Measure Of Quality In Science Measuring What's Measurable Rather Than What's Valid," *Scientometrics* 15.3-4 (1989): 189-203; Nederhof, A. J., and Zwaan, R. A. "Quality Judgments of Journals as Indicators of Research Performance in the Humanities and the Social and Behavioral Sciences," *Journal of the American Society for Information Science* 42.5 (1991): 332– 40; Nederhof Anton J., "Bibliometric Monitoring Of Research Performance In The Social Sciences And The Humanities: A Review," *Scientometrics* 66.1 (2006): 81-100; Leydesdorff Loet, "Can Scientific Journals Be Classified in Terms of Aggregated Journal-Journal Citation Relations Using the Journal Citation Reports?," *Journal Of The American Society For Information Science And Technology* (March 2006): 601-614.
5. Boyack, Kevin W., Klavans Richard, Börner Katy," Mapping The Backbone Of Science," *Scientometrics* 64.3 (2005): 351.374
6. Even though the end results of the “citation networks” give a scholar a good overview, they are still not integrated into the academic research facilities like Web of Science, Science Citation Index, Social Citation Index or Arts & Humanities Citation Index etc. The main reason is that once a citation network is derived, it becomes a static entity; it covers a limited time and scope. Thus citation networks, by their very definitions and aims, fail to keep up with new publications.
7. You can find more information about Pajek, its history and application areas at `<http://vlado.fmf.uni-lj.si/pub/networks/pajek/default.htm>`, last accessed 2006-11-15.
8. Please check `<http://www.w3.org/>` to see the latest standards of World Wide Web in relation to 3d publishing, last accessed 2006-11-15.

Gender, Race, and Nationality in Black Drama, 1850-2000: Mining Differences in Language Use in Authors and their Characters

Shlomo Argamon *(argamon@iit.edu)*
Linguistic Cognition Lab
Illinois Institute of Technology

Russell Horton *(russ@diderot.uchicago.edu)*
Digital Library Development Center
University of Chicago

Mark Olsen *(mark@barkov.uchicago.edu)*
ARTFL Project
University of Chicago

Sterling Stuart Stein *(stein@ir.lit.edu)*
Linguistic Cognition Lab
Illinois Institute of Technology

The Black Stage has been an important focus of evolving Black identity and self-representation, touching on many of the most contentious issues in American history since Emancipation—such as migration, exploitation, interracial unity, racial violence, and civil rights activism (Hill 2003). Alexander Street Press (ASP), in collaboration with the ARTFL Project, has developed an extensively tagged database of over 1,200 plays containing 13.3 million words by Black playwrights, from the middle of the 19th century to the present, including many previously unpublished works. Like other ASP datasets, the Black Drama database is remarkable for its detailed encoding and amount of metadata associated with authors, titles, acts/scenes, performances, and characters. Of particular interest for this study are the data available for authors and characters which are stored as "stand-off mark-up" data tables. The character table, for example, contains some 13,360 records with some 30 fields including name(s), race, age, gender, nationality, ethnicity, occupation, sexual orientation, performers, if a real person, and type. More extensive information is available for authors and titles. The character data are joined to each character speech, giving 562,000 objects that can be queried by the full range of character attributes. The ARTFL search system, PhiloLogic, allows for joining of object attribute searches, forming a matrix of author/title/character searching. For example, one can search for words in speeches by female, black, American characters depicted by male, non-American authors in comedies first published during the first half of the 20th century.

While user-initiated full-text searches on such author and character attributes can help answer specific questions, we believe that advanced text data mining systems have the potential to reveal important new patterns of variation in general language use, broken down by various combinations of author and character attributes. Initial work on racial epithets in this collection has revealed striking differences in the use of such language between male and female authors and characters, as well as American and non-American authors. While illustrative, such micro-studies can do no more than hint at larger discursive and representation issues that we believe can be identified by text mining techniques. Prior studies using text mining for analyzing variation in language use among different classes of authors have succeeded in identifying meaningful linguistic features distinguishing author gender, age, and personality type (e.g. Argamon et al. 2003, Koppel et al. 2002).

As might be expected of a collection of a particular class of literary texts, the Black Drama database cannot be considered a "random" sample. The database contains 963 works by 128 male playwrights and 243 pieces by 53 female playwrights. Plays by Americans dominate the collection (831 titles), with the remaining 375 titles representing the works of African and Caribbean authors. The database contains 317,000 speeches by 8,392 male characters and 192,000 speeches by 4,162 female characters. There are 336,000 speeches by 7,067 black characters and 55,000 by 1,834 white characters with a smattering of speeches by other racial groups. As would be expected, the predominance of American authors is reflected in the nationalities of speakers in the plays, with 272,000 speeches, compared with 71,000 by speakers represented as coming from a variety of African nations.

Using these data, we are examining the degree to which machine learning can isolate stylistic or content characteristics of authors and/or characters having particular attributes—gender, race, and nationality—and the degree to which pairs of author/character attributes interact. The first step to discover if lexical style- or content-markers can be found which can be used to reliably distinguish plays or speeches broken down by a particular characteristic, such as gender of character. A positive result would constitute strong evidence for distinctive, in this case, male and female, character voices in the sample of plays. If distinctiveness can be shown, we then seek some 'characterization' of the differences found, in terms of well-defined grammatical or semantic classes. The experimental protocol which we have been developing for this purpose, as applied by, e.g., Argamon et al. (2003), addresses both goals using techniques from machine learning, supplemented by more traditional computer-assisted text analysis.

First, to analyze a corpus of texts for distinctiveness, we need to determine if effective predictive models can be learned from the texts, which accurately classify new texts (that the system has not seen). The standard technique of 10-fold cross-validation can be used to estimate the usefulness of a learning method for constructing models that work for 'out-of-sample' data. The corpus is divided into 10 random subsets, and training is repeated for each of the 10 sets of 9 of those subsets, with accuracy of the resulting model is measures on the remaining subset. The average of these 10 numbers then forms a reasonably good estimate of how a model learned on the entire corpus would perform for new data. If this cross-validation accuracy is high (at least 70% for 50-50 balanced data), we may conclude that the two classes of texts in the corpus are linguistically 'distinctive'.

Second, to characterize the difference found (if any), all textual features extracted (frequencies of lexemes, lemmas, parts-of-speech, etc.) are ranked by some measure of how each of them enables prediction of a text's correct class. One such measure is to use feature weights computed during learning (for SVM or Naive Bayes learners, which compute such weights). Features with high weights (positive or negative) are the most influential in classifying a text as one class or the other (dependent on the sign of the weight). While a direct measure of influence, these weights can be difficult to interpret since the effect of a feature must be considered in the context of all the other features influencing classification. Another approach is to use a function that measures the 'distinguishability' of a feature without regard to other features, such as information gain or binormal separation (Forman et al. 2003). The downside here is that some features may be of little use alone, but in conjunction with others may have great discriminating power. In the current study we will be examining the usefulness of multiple measures of both types, and see which approach proves to be the more useful.

We conducted preliminary tests using the SVM-Light system (Joachims 1999) with PGPDT (Zanghirati 2004, Zanni 2006) to build the models. We extracted all speeches with character gender attributes from the corpus, splitting them into tokenized word frequency vectors for all authors, all characters, male authors, female authors, male characters, and female characters. For each of these, PGPDT built a model to identify authors and characters by gender.

Table One: Accuracy over 10-fold cross-validation						
Given:	**Female Author**	**Female Speaker**	**Male Author**	**Male Speaker**	**Full Sample**	**Full Sample**
Find gender of:	***Speaker***	***Author***	***Speaker***	***Author***	***Speaker***	***Author***
Accuracy	0.697	0.831	0.789	0.886	0.774	0.882
Majority Class	0.545	0.745	0.694	0.847	0.666	0.813

Table One

As indicated in Table One, the system the system correctly identified 88.2% of the authors' gender and 77.4% of the speakers' gender. Performance varied when examining subsets of the corpus, from 86.6% for gender of author in male characters to 69.7% for gender of speaker in female authors. All of these indicators show significant differences in words used by male and female authors and speakers. The differences in accuracy in male and female author/character may, however, result from the fact that male authors tend to include fewer female characters and that the have fewer words as well as having fewer female authors in the sample. This is shown in the Majority row of Table One, which indicates the rate of male instances for each assessment. For female authors, male characters constitute 54.5% of the speakers.

We then equalized a test sample for class by discarding instances until we have a balanced set with an attempt to correct for word frequencies as well. As shown in Table Two, author and character gender can be discriminated well. Furthermore, we see that identification of author gender is consistently more accurate than gender of speaker.

Table Two: Accuracy over 10-fold cross-validation						
Name	**Female Author**	**Female Speaker**	**Male Author**	**Male Speaker**	**Author**	**Speaker**
Given:	**Female Author**	**Female Speaker**	**Male Author**	**Male Speaker**		
Find gender of:	***Character***	***Author***	***Character***	***Author***	***Character***	***Author***
Accuracy	0.719	0.828	0.803	0.837	0.799	0.864
Average words/document						
Male	770.1	877.8	853.4	645.2	859.4	750.8
Female	877.9	877.9	853.0	645.0	859.4	750.8

Table Two

As shown in Table 3, male authors/speakers are correctly somewhat more often in five of the six cases, with the sole exception of almost exactly the same correct identification of speaker gender in female authors.

Table Three: Accuracy of Gender Identification by Category						
Name	**Female Author**	**Female Speaker**	**Male Author**	**Male Speaker**	**Author**	**Speaker**
Given:	**Female Author**	**Female Speaker**	**Male Author**	**Male Speaker**		
Find gender of:	***Character***	***Author***	***Character***	***Author***	***Character***	***Author***
Male Accuracy	0.718	0.857	0.817	0.866	0.811	0.889
Female Accuracy	0.721	0.799	0.788	0.808	0.786	0.839

Table Three

These results suggest that female lexical choices, both used by authors and depicted in characters, are somewhat less marked (or more varied) than male use of language. The full paper will explore this phenomenon in greater detail. While the accuracy of identification of gender in significant in all the cases we example, we expect that using some sort of feature set selection, as in, for example, Hota, Argamon, and Chung (2006), will improve the precision of the identification.

From this preliminary analysis, it would appear that the authorial gender is rather more readily identified than the represented gender in characters. Initial examination of the features that best predict gender of character, as identified by information gain, range from the expected (male characters speak of wives and swear more frequently), to the somewhat

opaque, such as the words 'nonsense' and 'reason' being strongly male associated. It is also important to note that both strongly male and female character terms in this sample are used at about the same rates (per 10000 words) by male and female authors. This suggests that male and female authors are able to use certain linguistic gender markers effectively. As noted, it is significant that the machine learning algorithms employed are less accurate in most cases in identifying female as opposed to male authors and speakers.

The final paper will report results using similar techniques to examine the the degree to which additional character attributes -- race and nationality -- can be distinguished and, if this proves to be effective, examine the most important features distinguishing between the language of white and black speakers and American/non-American speakers. An initial examination of racial slurs in this dataset suggests that speaker race and nationality may also be readily identified.

As we have seen, machine learning and text mining techniques can support higher orders of generalization and characterization than the more traditional user-driven search methods widely used in computer-aided textual research. This approach is most effective when used in relatively constrained experiments where classification criteria are clearly defined, such as the social attributes of authors and their characters. Some results may be trivial on a literary level— of course men talk more of wives and only women tend to call other women hussies —but such common sense results allow us to argue that the technique gives meaningful results, and so odd results should be examined further, using more traditional systems like PhiloLogic. We therefore argue that approaching the interpretative process starting with highly structured and constrained experimental hypotheses, we can take advantage of machine learning methods to find new and unexpected foci for examining literary questions, which may in turn shed new light on critical issues such as race and gender.

Bibliography

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. "Gender, Genre, and Writing Style in Formal Written Texts." *Text* 23.3 (2003).

Forman, George, Isabelle Guyonl, and André Elisseff. "An Extensive Empirical Study of Feature Selection Metrics for Text Classification." *Journal of Machine Learning Research* 3.7-8 (2003): 1289-1305.

Hill, Errol G., and James V. Hatch. *A History of African American Theatre.* New York: Cambridge University Press, 2003.

Hota, Sobhan, Shlomo Argamon, Moshe Koppel, and Iris Zigdon. "Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters." Paper presented at Digital Humanities 2006, Paris Sorbonne, 5-9 July 2006 . 2006.

Hota, Sobhan, Shlomo Argamon, and Rebecca Chung. "Gender in Shakespeare: Automatic Stylistics & Gender Character Classification Using Syntactic, Lexical and Lemma Features." Paper presented at the Chicago Colloquium on Digital Humanities and Computer Science, Nov. 2006, Chicago, Illinois. 2006.

Joachims, Thorsten. "Making large-Scale SVM Learning Practical." *Advances in Kernel Methods - Support Vector Learning.* Ed. Bernhard Schölkopf , Christopher J. C. Burges and Alexander J. Smola. Cambridge, MA: MIT Press, 1999.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. "Automatically Categorizing Written Texts by Author Gender." *Literarcy & Linguistic Computing* 17.4 (2002): 401-12.

Olsen, Mark. "Gender Representation and Histoire des Mentalités: Language and Power in the Trésor de la Langue Française,." *Histoire et Measure* VI (1991): 349-73.

Olsen, Mark. "Écriture Féminine: Searching for an Indefinable Practice?" *Literary & Linguistic Computing* 20 Supplement 1 (2005): 147-164.

Olsen, Mark. "Making Space: Women's Writing in France, 1600-1950." Paper presented at ALLC/ACH 2004 Conference, Göteborg, Sweden. 2006.

Zanghirati, Gaetano, and Luca Zanni. "A Parallel Solver for Large Quadratic Programs in Training Support Vector Machines." *Parallel Computer* 29 (2003): 535-551.

Zanni, Luca, Thomas Serafini, and Gaetano Zanghirati. "Parallel Software for Training Large Scale Support Vector Machines on Multiprocessor Systems." *JMLR* 7 (2006): 1467-1492.

Software Sites

- PhiloLogic: `<http://philologic.uchicago.edu/>`
- Parallel GPDT: `<http://www.dm.unife.it/gpdt/>`
- SVM-Light: `<http://svmlight.joachims.org/>`

Discourse, Power and *Écriture Féminine*: Text Mining Gender Difference in 18th and 19th Century French Literature.

Shlomo Argamon *(argamon@iit.edu)*
Linguistic Cognition Lab
Illinois Insitute of Technology
Jean-Baptiste Goulain *(jibai31@gmail.com)*
Linguistic Cognition Lab
Illinois Insitute of Technology
Russell Horton *(russ@diderot.uchicago.edu)*
Digital Library Development Center
University of Chicago
Mark Olsen *(mark@barkov.uchicago.edu)*
ARTFL Project
University of Chicago

It is well documented that men and women use informal language, such as conversation and correspondence, in rather different ways, reflecting a wide variety of cultural forces and practices (Tannen 1990, Eckert & McConnell-Ginet 2003). Recent work has suggested that gender differences may also be found in more formal, written communications. Koppel, Argamon, et al (2002) have shown that gender of author can be accurately predicted between 70 and 80 percent of cases of published samples from the British National Corpus, using machine learning and text mining techniques. Using simple statistical and collocation techniques, Olsen (2005) has argued that there are distinct gender differences in literary French from the 17th to the early 20th centuries. He (Olsen: 2004) further proposed that the differences between male and female writing during this time does not appear to support a "two cultures" model (Liberman 2004), but that it was a more conscious political activity of domain specific writing in which men and women deployed the same language.

We are addressing several issues that arise from our previous work on gender and public writing in this study. The first is to apply machine learning and data mining techniques to the same sample used by Olsen to examine the degree to which conclusions found in Koppel, Argamon, et al. can be replicated for larger sample of French literary texts. The methodologies used by Koppel, Argamon, et al. and Olsen may have privileged identification of particular features, with Koppel focusing on more stylistic aspects of writing and Olsen examining significant frequency differences of content words. Use of a common approach will allow us to examine the degree to which gender differences can be compared between cultures and time periods. We also expect use of machine learning techniques will allow us to test Olsen's claim that gender differences in French literature are best explained as examples of what Liberman describes as "dominance" theories of gender distinction. If the features that best characterize gender difference are more "stylistic", this would suggest that "difference" or two culture models are more applicable, but if differences are shown in content words and these words are used in similar ways, we would argue that the "dominance" or political theory may be a more effective explanatory model. And finally, we are comparing several distinct statistical techniques, most notably Support Vector Machine (SVM) and information gain models, to see which are most effective at extracting weighted features that can be used to interpret observed differences in male and female writing, comparing these with more traditional examinations of differences in term frequency. Becuase of the need to pick apart the learned model to examine its inner workings, with the model itself as the end product of the learning process, some Digital Humanities applications of machine learning techniques will differ in emphasis from more common data mining tasks, where the end goal is often to maximize performance on the classification of unknown data, and the model may be viewed as a black box.

This study is based on the creation of two samples of 300 texts roughly balanced by genre, collection and time period driven by texts by French women writers available to us. For each of the 300 texts by 67 female authors (18.5 million words), we selected the chronologically closest male document by genre and, where available, by collection, leading to a comparison collection of 300 texts by 170 authors (27 million words). As noted in Olsen (2005), the samples are largely drawn from the 18th-early 20th centuries with strongest representation in the 19th century, owing to the predominance of romantic novelists in the available collections of female writers. The sample is also skewed by the presence of many works by particular notable authors, in particular George Sand. The ARTFL project is currently adding a large block of texts to our collection of French Women Writers (FWW) and these will be integrated into the full study, depending on how much is available in time.

Each text in the corpus has been prepared by first tokenizing it and running it through TreeTagger to determine lemmas and parts-of-speech for each token. In cases of uncertainty, each possible assignment is assigned a probability value between 0 and 1 by TreeTagger; currently, we only consider the most likely analysis for each token (though we will apply more sophisticated techniques in the sequel). We consider in the

study several types of lexical features by which to represent the texts numerically: Word Frequencies (WF), Lemma Frequencies (LF), Part-of-speech Frequencies (PF), Word Bigram Frequencies (WBF), Lemma Bigram Frequencies (LBF), and POS Bigram Frequencies (PBF). For each such feature set, we compute a vector of numeric values for each text, where elements of the vector represent the relative frequencies of various features in the set. For example, a WF vector may contain entries for "la", "l'", "femme", and "femmes", while the corresponding LF vector will contain entries just for "la" and "femme", and the corresponding LBF vector may contain an entry for "la/femme". Features that occur less than 10 times in the entire corpus are elided; in some experiments, more stringent criteria for feature inclusion will be used, to see how few features we can get away with.

For classification, we are applying the SVM method; SVMs are one of the most successful modern methods for text categorization (Joachims 2002). We are currently working with the Weka (Witten & Frank 2005) implementation, using a linear kernel and the default parameters. (Other options do not appear to improve accuracy by much, so we used the simplest option, which also enables easier interpretation of the results.) As the work progresses, we will be evaluating computational efficiency, and may switch to other, faster, software codes if necessary. The classification model constructed is a "linear model", which means that it assigns a single numeric weight to each input feature, positive for one input class (say, "femme") and negative for the other (say, "homme"). The magnitude of the weight corresponds to the importance or influence of the feature's value for classification; high weights indicate those features with the most effect on the classification of a given text (in a specific model). Once classified by the SVM method, we apply the information gain and other metrics (Forman et al. 2003) to identify those features that are most relevant to the classification task. Tokens and lemmas found in this step will be compared to differential frequency tables already used by Olsen.

Preliminary classification results using Weka Sequential Minimal Optimization (SMO) implementation of a support vector classifier with 10 fold cross-validation confirm that author gender in this sample can be detected with surprising reliability. For the entire sample, gender of author was correctly identified 85.9% of the time using word lemmas and 85.7% on tokens, with lower performance for both generic POS (73%) and more specific POS (75%). A second test on paired 100 document male/female collections, to reduce the number of works by Sand and smooth other main sample anomalies, achieved slightly better performance, with correct classification by author gender at 87% for both surface word forms (tokens) and lemmas, with POS again performing less well at 76.5% and POSgroups (abstract POS) at 72%. The accuracy performance of classifications based on tokens and lemmas (86% and 87%) is somewhat higher than the 70-80% performance for a sample modern English documents reported by Koppel et al. (2002) This may be due to our use of all of the words in this experiment, suggesting that men and women tend to write about different things and use somewhat different vocabularies.

Table 1 shows the confusion matrix from the Weka SMO classification expressed as the correct classification rate broken down by feature type. In both samples, the highest performing features (tokens and lemmas) are considerably better at identification of male authors than female authors. We have noted this difference in other recent work on modern English data. Surprisingly, the opposite is true (in 3 of 4 instances) for part of speech features, though it is unlikely that the differences are statistically significant.

Table 1: Accuracy over 10-fold cross-validation

Feature	Token	Lemma	PoS	PoSgroup
2x300 document sample				
Male	0.883	0.873	0.730	0.697
Female	0.833	0.844	0.757	0.787
All	0.857	0.859	0.744	0.742
2x92 document sample				
Male	0.913	0.924	0.739	0.739
Female	0.815	0.815	0.783	0.696
All	0.864	0.870	0.761	0.717

Table 1

Using a variety of techniques, we can effectively distinguish between male and female authors. Examination of the highest weighted features using the information gain (IG) measure on the 2x92 subset sample reveals clear agreement with previous work. Female authors show a strong preference for writing about women (noted by the pronoun *elle*), adopt a more personal and reflective frame (je, me), and address (vous). This is consistently shown in both selected tokens and lemmas, and is also indicated in the part of speech analysis by a preference for the use of personal pronouns, indefinite pronouns and possessive pronouns. Examination of tokens assigned a high information gain in the subset selected to correct for sample biases reveals the same female concern for descriptions of internal, subjective and emotive states described by Olsen (2005). These terms include émotions, amitié, chagrin, courage, craint* esprit* desire. generosité, larmes, motifs, peines, penser, plaisirs, réflexions, sensible, sentiment, sentis, soin*, and souffrance. The IG measures are notable for some striking absences from Olsen's examination, including terms like amour* and aim* which were strongly correlated to female discourse, as well as many kinship terms (oncle, maman, enfant, parents, etc.). Terms with high information gain values that are more common in male than female writers are less clearly grouped and are

missing the strong preference noted in Olsen for abstractions and numbers. However, when examining the PoS measures, ordinal numbers are preferred by male writers. These variations arise from using the IG measure as opposed to differences in relative frequencies to assess "importance" and examination of a small but more random sample, issues which will be explored in the full paper.

Text mining techniques can identify gender of author using a variety of models with impressive accuracy. The features ranked most highly in learned classification tend to confirm previous results and point to the possibility of cross-language patterns of gendered language use. The full paper will explore the reasons for the higher accuracy performance in identifying male authors and examine in more detail the male and female author feature sets with specific reference to the differences in ranked features arising from different techniques. Finally, we will address the larger question of whether the clear distinctions of male and female public writing in this sample arise from a "two cultures" interpretation or more conscious political stances by a systematic analysis of word collocations around particularly gender specific terms.

Software Sites

- PhiloLogic:<http://philologic.uchicago.edu/>
- TreeTagger:<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>
- Weka 3: <http://www.cs.waikato.ac.nz/ml/weka/>

Bibliography

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. "Gender, Genre and Writing Style in Formal Written Texts." *Text* 23.3 (2003): 321–346.

Eckert, Penelope, and Sally McConnell-Ginet. *Language and Gender*. Cambridge: Cambridge University Press, 2003.

Forman, George, Isabelle Guyon, and André Elisseff. "An Extensive Empirical Study of Feature Selection Metrics for Text Classification." *Journal of Machine Learning Research* 3.7-8 (2003): 1289-1305.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. "Automatically Categorizing Written Texts by Author Gender." *Literary &Linguistic Computing* 17.4 (2002): 401-12.

Liberman, Mark. "Sexing Rivka." *Language Log*. May 8, 2004. <http://itre.cis.upenn.edu/~myl/languagelog/>

Liberman, Mark. "Gender and Tags." *Language Log*. May 9, 2004. <http://itre.cis.upenn.edu/~myl/languagelog/>

Mitchell, Tom. *Machine Learning*. McGraw Hill, 1997.

Olsen, Mark. "Making Space: Women's Writing in France, 1600-1950." Paper presented at ALLC/ACH 2004 Conference, Göteborg, Sweden.

Olsen, Mark. "Gender Representation and histoire des mentalités:Language and Power in the Trésor de la langue française." *Histoire et measure VI (*. 1991. 349-73.

Olsen, Mark. "Écriture Féminine: Searching for an Indefinable Practice?" *Literary & Linguistic Computing* 20 (2005): 147-164.

Tannen, Deborah. *You Just Don't Understand*. William Morrow, 1990.

Witten, Ian H., and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Technique*. 2nd. Morgan Kaufmann, 2005.

Viewing Texts: An Art-Centered Representation of Picasso's Writings

Neal Audenaert *(neal@csdl.tamu.edul)*
Texas A&M University
Unmil Karadkar *(unmil@csdi.tamu.edu)*
Texas A&M University
Enrique Mallen *(e-mallen@tamu.edu)*
Texas A&M University
Richard Furuta *(furuta@cs.tamu.edu)*
Texas A&M University
Sarah Tonner *(sarahtonner1@yahoo.com)*
Texas A&M University

Introduction

The *Picasso Project* is a web-based, dynamic *catalogue raisonné* currently containing digital images and descriptive metadata for more than 11,000 of Picasso's works (Mallen, 2006). The catalogue includes commentary for many artworks, along with notes detailing sales of the item, exhibitions in which it has been displayed, and scholarly literature in which it has been cited. In addition to these works, the catalogue also provides extensive biographical information including nearly 7,500 entries. The biography is linked to artworks, photographs of key people, maps of the various places Picasso lived and worked, and related documents (for example, Picasso's birth certificate). Tools are provided to support side-by-side comparisons of two artworks and to build custom subcollections from which an illustrated, color catalogue can be automatically generated for printing. Within this context, we have become increasingly interested in exploring Picasso's writings which have, until recently, been largely neglected in favor of his more well known work in painting, sculpture, and collage.

Picasso, first and foremost an artist, wrote texts that are strikingly visual, both in terms of the design of the texts themselves, as well as the decorative elements that adorn the pages on which these texts are found. These intensely visual writings present his editor with significant challenges—challenges that exemplify the TEI consortium's guidelines for when TEI should not be used (Lavagnino, 2006). As Lavagnino points out, "to make the [TEI] edition work as intended it is generally necessary to interpret features and not merely reproduce their appearance." Picasso's writings do not readily yield to a single fixed interpretation that can be understood by an editor and transcribed in some definitive form. Indeed, their interest stems, in part, from their complex and indeterminate nature. Furthermore, transcription involves the selection of relevant features of a work and production of a digital (text plus encoding) description of them. In cases where this digital description has little value for analysis, alternative forms of presentation must be pursued. The two examples of this cited by Lavagnino, "works intended as mixtures of words and images, and very complex draft manuscripts in which the sequence of text or inscription is difficult to make out," are both characteristic of Picasso's writings.

Despite these difficulties, there have been a number of efforts to produce transcriptions of Picasso's writings as published books (Bernadac, 1998; Michaël, 2005). This approach gives primacy to the textual content of these works, at the expense of losing almost all of the visual elements. To remedy this, most editions include facsimile reproductions of illustrative samples of his writings alongside the transcriptions. While this is helpful in conveying a sense of the original context of these works, it remains inadequate for many purposes. Beyond limitations of scale (it is only feasible to include a limited number of facsimiles), this approach treats the literal textual elements of Picasso's writings as

Figure 1: Two of Picasso's intensely visual texts.

primary, bringing in the images of the original, visually constructed pages in order to illustrate and elaborate. This approach to remediate the writings of Picasso divorces Picasso the writer from Picasso the artist, limiting the productive interchange of ideas that might result from a blending of literary and art history-based approaches.

While Picasso's writings provide a compelling example of the limitations of a purely textual approach to representing manuscripts and other documents, this problem is not unique.

Within the digital textual studies community there is an increasing recognition of the need to pair robust text encoding with access to images of the original source material (Dicks, 1997; McGann, 2001). Over the past decade, a number of projects have focused on presenting documents as images while providing additional support via transcriptions (McGann, 1996; Viscomi, 2002). Others have supplemented textually-oriented systems by providing access to digitized images of the original documents in a variety of formats (Furuta, 2001; Robinson, 1996). Image based representations of documents have placed particular strain on hierarchical methods for representing and encoding the features of a text and alternative formal models have been proposed (Dekhtyar, 2006; Renear, Durand, and Mylonas, 1993).

Approach

Within the *Picasso Project*, we have encountered the problem of developing digital representations of texts from a different perspective than that found in either print based transcriptions or in digital textual projects—namely, we have treated Picasso's writings, first as works of art, and secondarily as art that contains text. By conceiving of and contextualizing the writings of Picasso in a form common to traditional approaches to art (that is, the *catalogue raisonné*) we are able to reap immediate benefits for understanding these texts in ways not readily supported by the tools grounded in a textual approach to these works. This approach to visualizing Picasso's texts places them squarely within the context of the other works that he was painting and thinking about at the same time. This helps inform our understanding of both the artworks that Picasso produced as well as his writings. For example, Figure 2 shows a typical thumbnail view of the last fifteen items in the catalogue for the year 1935. In the poem in the center of the last line (OPP.35:004), Picasso refers to a small girl. In the context of drawings made by Picasso around the same time we see connections with Picasso's daughter, Maya, at three and three and a half months old (OPP.35:031 and OPP.35:032), followed by Marie-Thérèse, Maya's mother (OPP.35:034). Similarly, using the comparison tool to compare a text dated 28 November 1935 with a painting made earlier that year highlights possible connections between verbal imagery of the text and the visual imagery of the painting. References in the text to "tongue of fire," "stabbing," and "the eye of the bull" take on new meanings when seen in this context.

In addition to visually contextualizing writings in relationship to Picasso's other works, the digital *catalogue raisonné* (unlike a corresponding print version) allows us to make accessible images of Picasso's writings suitable for reading and analysis. Closer examinations of the text enables scholars to consider multiple states of a text, to see annotations, deletions, and additions to a text, to explore Picasso's use of color to provide structural divisions or graphical bars rather than traditional punctuation to divide conceptual segments of the text. These tasks, which are difficult or impossible to perform

Figure 2: la petite fille" with drawings of Maya and Marie-Thérèse

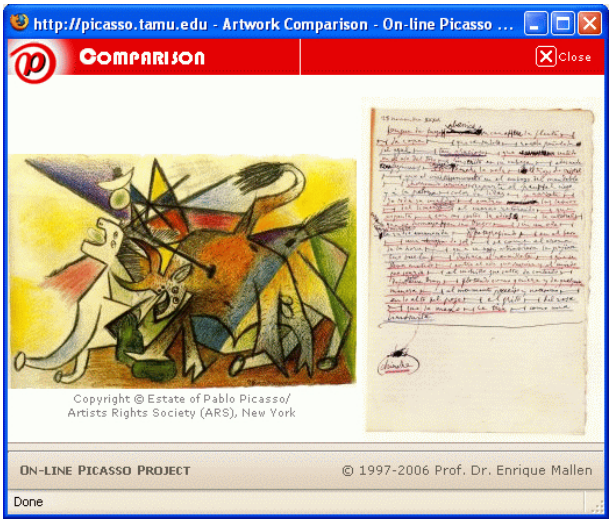


Figure 3: Courses de taureaux and "lengua de fuego"

using transcriptions alone, are encouraged by the online presentation. To further enable access to the textual component of these works, we are initially adding transcriptions of these texts to the biographical section of the catalogue. Since the online catalogue presents the biographical text in parallel with the artworks, this technique permits easy cross-referencing between the transcriptions and images while we investigate more sophisticated means for encoding and presenting the content of these writings.

Implications and Future Work

Framing our approach to Picasso's writings in terms of artworks that contain text, encourages us to look for text in artworks in general. Like many other artists of his day, Picasso began incorporating words into his artworks in a variety of ways and forms, notably in the newspaper clippings pasted

into his *papiers-collés*. These works reinforce our conviction that we need develop tools for working with text in art grounded to the needs of the artistic disciplines rather than those of the traditional textual studies community.

Picasso's unique works offer fertile ground for exploring the techniques and tools that can be applied to visually constructed texts but much work is needed—both in terms of understanding the needs of the scholars and other readers interested in Picasso's writings as well as formulating new models for representing and working with these texts in a digital environment. We are currently investigating the potential for techniques based in spatial hypertext research for interpreting and presenting these texts. In spatial hypertext, Figure 3: Courses de taureaux and "lengua de fuego" nodes of information that are connected by visual elements (for example, text style, 2-D position, color) rather than by explicit links (Shipman, 1999). Our current efforts are focused in understanding how the texts might be sub-divided into their constituent parts and manipulated in a 2-D space in ways that enhance understanding. We are also looking at how formal relationships between these parts can be incrementally added as an expression of an editor/reader's evolving understanding. In addition to purely image based approaches, we are interested in studying methods of encoding the textual content of Picasso's works to support content based retrieval, enable automatic processing, and facilitate reading.

Summary

By translating a traditional print-based approach for working with a large corpus of artworks, the *catalogue raisonné*, into a digital format, we increase the level of support provided for three scholarly primitives (Unsworth, 2000): comparing (either two works side by side, or many works in a thumbnail view), sampling select artworks from the collection as a whole, and representing the artworks not merely as thumbnails, but also with higher resolution images. With these enhancements, the digital *catalogue raisonné*, though not its printed counter-part, provides a natural medium for presenting the writings of Picasso. In this context, his writings are presented against the backdrop of other artworks while enabling the textual elements of these artworks to be read and carefully studied as texts. In addition to the immediate benefits that this approach brings in terms of accessing Picasso's writings, it also offers a new paradigm for working with these texts that suggests several promising directions for further work.

Bibliography

Bernadac, M. L., ed. *Picasso: Collected Writings*. London: Aurum Press Ltd, 1989.

Dekhtyar, Alex, et al. "Support for XML Markup of Image-based Electronic Editions." *International Journal on Digital Libraries* 6.1 (2006): 55-69.

Dicks, R. S. "Third Commentary on "What is Text Really?"." *SIGDOC Asterisk J. Computer Documentation* 21.3 (1997): 36-39.

Furuta, Richard, et al. "The Cervantes Project: Steps to a Customizable and Interlinked On-Line Electronic Variorum Edition Supporting Scholarship." Ed. P. Constantopoulos and I. Sølvberg . ECDL 2001, Lecture Notes In Computer Science. Heidelberg: Springer-Verlag, 2006.

Lavagnino, John. "When Not to Use TEI." *Electronic Textual Editing*. Ed. Lou Burnard, Katherine O'Brien O'Keeffe and John Unsworth. New York: Modern Language Association of America, 2006. 334-338. Accessed 2006-11-10. <http://www.tei-c.org/Activities/ETE/Preview/lavagnino.xml>

McGann, Jerome. "Rethinking Textuality." *Radiant Textuality*. New York: Palgrave-Macmillian, 2001. Accessed 11-6-2006. <http://www.iath.virginia.edu/~jjm2f/old/jj2000aweb.html >

McGann, Jerome. "The Rosetti Archive and Image-Based Electronic Editing." *The Literary Text in the Digital Age*. Ed. Richard J. Finneran. Ann Arbor: University of Michigan Press, 1996.

Michaël, A., ed. *Picasso: Poèmes*. Paris: Le Cherche Midi, 2005.

Renear, Allen, David Durand, and Elli Mylonas. *Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies* . Accessed 2006-11-01. <http://www.stg.brown.edu/resources/stg/monographs/ohco.html >

Robinson, P. M. W. "Is There a Text in These Variants'." *The Literary Text in the Digital Age*. Ed. Richard J. Finneran. Ann Arbor: University of Michigan Press, 1996.

Shipman, F. M., and C. C. Marshall. "Spatial Hypertext: An Alternative to Navigational and Semantic Links." *ACM Computing Surveys* 31.4 (1999). <http://doi.acm.org/10.1145/345966.346001>

Mallen, Enrique, ed. *The Picasso Project* . Texas A&M University . Accessed 2006-11-06. <http://picasso.tamu.edu/ >

Unsworth, John. "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" Paper presented at Humanities Computing: Formal Methods, Experimental Practice, London, May 13, 2000. 2003. Accessed 11-6-2006. <http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html >

Viscomi, J. "Digital Facsimiles: Reading the William Blake Archive." *Image-based Humanities Computing* 36.1 (2002): 27-48.

A Flexible System for Text Analysis with Semantic Networks

Loretta Auvil *(lauvil@ncsa.uiuc.edu)*

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Eugene Grois *(egrois@gmail.com)*

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Xavier Llorà *(xllora@illigal.ge.uiuc.eduu)*

University of Illinois at Urbana-Champaign

Greg Pape *(gpape@gpape.com)*

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Vered Goren *(vered@ncsa.uiuc.edu)*

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Barry Sanders

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Bernie Acs

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

1. Introduction

The explosive growth of digital and digitized text creates opportunities for scholars and students to conduct new analyses and develop unique insights about our written culture and heritage. To effectively use large collections of textual data, scholars and students need flexible, easy to use tools that provide powerful analysis and visualization.

1.1 Goals

The Automated Learning Group is developing interactive tools for mining large, complex semantic networks, which were automatically extracted from a text corpus. The corpus could be a collection of documents or a stream of messages. The

semantic network represents the concepts in the collection as entities and relations. The visual environment allows the user to query the semantic network to retrieve portions of the graph. The display allows the user to view and navigate these networks to discover patterns in the collection.

1.2 Technical Approach

Our approach to the visual analysis of text documents implements extraction and visualization of important entities and the relations among them. This is accomplished in the following high-level steps:

1. Perform lexical analysis of textual document set
2. Extract key entities and relations from the text
3. Compose entity-relation triples into semantic network model summarizing key information in document corpus
4. "Publish" the semantic network in a repository
5. Visualize semantic network model as graph
6. Search and prune visual graph through appropriately structured queries against semantic network model.

2 Text Analysis

Textual processing is accomplished by passing documents through a series of syntactic, semantic, and functional analysis tools. These tools primarily consist of D2K [1]and T2K, our own analytic software, working in concert with GATE [2]and MontyLingua [3]. These tools are seamlessly connected together into the D2K visual programming environment

The GATE toolkit is utilized for standard syntactic processes including tokenization, sentence splitting, and part-of-speech tagging. Syntactically annotated documents are passed to GATE's Named Entity (NE) tagger. The GATE NE tagger identifies proper nouns in the text that belong to relevant categories such as "Person", "Organization", and "Location." NE tagging passes through three modules that perform different kinds of co-reference[2]. The final output of the NE tagging and co-referencing processes are a set of annotations that identify references to "Person", "Organization", and "Location" entities.

After NE extraction, we identify key relations in the text using a tool within MontyLingua. This Jister tool extracts sentence structures called "jists," which carry thematic-role information such as verb, subject, and object. This is similar to semantic role labeling [2], but less formally structured. An example jist:

Tiger Woods wrapped up the tournament at noon.
(Verb: 'wrap up', Subj: 'Tiger Woods',
Obj1: 'tournament',Obj2: 'at noon')

The next step involves the normalization of references in the MontyLingua jists with entities identified in the analysis. For example, "Tiger Woods" may have been identified as a entity:PERSON, and "tournament" referenced back to "Masters Golf Tournament" appearing earlier in the text. Tagging the verb "wrap up" as an entity:ACTION, and "at noon" as a TIME relation to the verb, we can generate triples as follows: `<Tiger Woods> <is-a> <entity:PERSON> <Masters Golf Tournament> <is-a> <entity:EVENT> <wrap up> <is-a> <entity:ACTION> <Tiger Woods> <actor> <wrap up>`

Figure 1 shows the D2K toolkit with a view of the text processing itinerary. The itinerary is a dataflow graph: the nodes in the figure are D2K modules (major processing blocks), connected by edges representing the flow of data during execution. The itinerary can exploit both data parallelism (multiple documents at the same time) and task parallelism (different modules in parallel). The D2K itinerary can be run on a desktop or scaled up to multiple computers or High Performance Computer systems.Figure 1 shows the D2K toolkit with a view of the text processing itinerary. The itinerary is a dataflow graph: the nodes in the figure are D2K modules (major processing blocks), connected by edges representing the flow of data during execution. The itinerary can exploit both data parallelism (multiple documents at the same time) and task parallelism (different modules in parallel). The D2K itinerary can be run on a desktop or scaled up to multiple computers or High Performance Computer systems.

3 Semantic Network Storage and Retrieval

The triples generated from the semantic extraction process are stored in an RDF [5]metadata store. We use Kowari, developed by Tucana Technologies [6]. Additional triples are generated to represent metadata in conformance with a common vocabulary, and user annotations can be included as well. Queries are coded in an SQL-like query language called iTQL. The result is a set of triples, which represents a semantic graph [6].

This architecture demonstrates a key design principle for robust Cyberinfrastructure. The analysis is decoupled from the visualization, so that a large scale analysis can asynchronously update the triples as new results are computed, while interactive tools will automatically pick up the new data by refreshing the query. The triples generated from the semantic extraction process can be combined with many other similar relation triples from many sources, and additional triples are generated to represent a common vocabulary.

4 Visual Investigation of Semantic Networks

Using the visual environment, investigators can perform searches over the semantic networks extracted from a text corpus. The familiar web browser paradigm was employed in the user interface design. The user interface allows one to construct more complex queries by incorporating multiple rules and filters. Each user query is converted into iTQL and executed by the Kowari server[6] . The query history is available as a pull-down menu, just as query histories are in a web browser. Investigators can directly observe semantic relationships between entities in an interactive link-node graph visualization. (Figure 2)

Relations between entities in the resulting semantic network graph are displayed as a link-node graph visualized using Prefuse [7], and also as a hierarchical tree of entities conforming to the common vocabulary. The subject and object entities in the relations are displayed as nodes in the graph visualization, predicates are displayed as links. This simplification of the more complex semantic network, stored in Kowari, provides a compact and usable abstraction of the important relations extracted from the text streams.

The Entity pane at the upper left displays lists of entities. The Relations pane at the lower left displays a list of relations, with additional options to highlight synonyms, antonyms, hyponyms or hypernyms based on WordNet. Selecting an entity in the left pane also highlights the corresponding node or edge in the visualization.

Every entity in the graph maintains a link back to the original text document from which it was extracted. By right-clicking on a node, and selecting View Source Documents, the text of the original document will be retrieved from the repository and displayed.

5 Collections

This tool can be applied for many different types of text, across one or many collections. In addition, it can analyze evolving collections of text, such as documents from one or more RSS feed.

This tool has been used as part of the Nora project, a multi-institution collaboration to produce software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries.[8]

For example, we are experimenting with the digitized text of the novel *Uncle Tom's Cabin*, available as part of the Early American Fiction Collection from the University of Virginia[9]. The system performs feature extraction in order to determine shared characteristics of the selected documents, such as chapters of the novel. The resulting arc-node graph can be viewed and navigated to discover patterns and relations within and among the texts. Figure 2 illustrates an example analysis of text from nineteenth century novels, showing the use of concepts related to "Mother" and "Man".[10]

6 Conclusion

This paper described interactive tools for mining large, complex semantic networks automatically extracted from a text corpus. These approaches can be applied to news feeds, technical literature, or literary collections.

We believe this is a promising approach, although we are only beginning to develop these tools for use by humanists, who will be the ultimate judges of the utility and validity of this approach. The general purpose semantic analysis is widely applicable to many types of text, though it is difficult to predict the impact of such analysis on our understanding of texts.

The query interface and graphical displays are still under development. The entity-relationship graphs may be quite complicated, so we must find new visual methods and metaphors to enable scholars and students to understand the information in the graphs, and to use them formulate hypotheses, and answer questions.

7 Acknowledgements

The Nora project is funded in part by the Mellon Foundation. This work was funded in part by the National Center for Advanced Secure Systems Research (NCASSR) at the University of Illinois at Urbana-Champaign (UIUC), a multi-institutional cybersecurity research team. NCASSR is led by the National Center for Supercomputing Applications (NCSA) and supported by funding from the Office of Naval Research (ONR). Substantial portions of the code were implemented by David Clutter and Fang Guo. Thanks to Patricia S. Taylor.

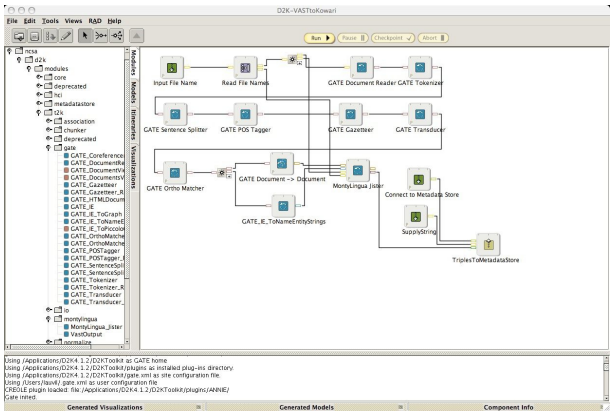


Figure 1: D2K Itinerary showing text processing.

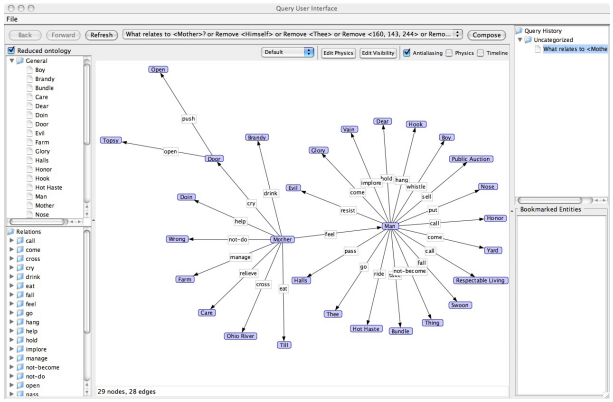


Figure 2: Visual interface showing result of a query.

TEI Constrained: Yet Another Presentation System

Syd Bauman *(Syd_Bauman@Brown.edu)*
Brown University Women Writers Project

The Text Encoding Initiative Consortium[1] claims that its new Guidelines (P5)[2] can be customized for almost any purpose, including writing new documents (as opposed to transcribing existing ones). This author regularly teaches workshops on text encoding using the Text Encoding Initiative Guidelines, and believes it is important that TEI advocates "eat our own dog food" — that is, we should use TEI to encode the documents used to teach TEI. This includes the slides displayed during a presentation or lecture, and any associated handouts. Initially the author's workshops were written in TEI Lite and the slides were created using Sebastian Rahtz's stylesheets for slides provided by the TEI Consortium. Quite quickly it became clear that a more reliable, better documented, and easier to use system was needed; the hallmark feature of such a system would be a highly constrained schema tailored to the needs of a workshop presentation.

Such a highly constrained schema tailored to the particular purpose has several important advantages. First, it makes authoring much easier. A large variety of elements that make no sense in the context of a workshop slide are available in TEI Lite, e.g. `<gap>` or `<interpGrp>`. When writing slides for the workshop, the author needs to pick and choose from among the available elements, including those which make no sense. Eliminating these makes the author's encoding task much simpler. Second, the semantics of elements can be expressed more explicitly. E.g., a `<slide>` element can be used instead of `<div type="slide">` . This may seem like a minor point, but the third advantage requires it: that constraints required for the purpose can be enforced in the schema. E.g., in TEI Lite the `<head>` element is an optional child of a `<div>`, whereas it can be made a required child of a `<slide>`, and still be an optional child of a `<div>`.

In order to ascertain how difficult it is to create such a highly constrained, specialized schema using P5, a new system was developed using the TEI's new ODD (one document does it all) system.[3] Because there are already quite a few presentation systems available to the workshop instructor, both proprietary and open, both XML-based and others, the new system is called Yet Another Presentation System, or "yaps". Accordingly, its icon is a dog. Yaps provides the capability to write a single

1. <http://alg.ncsa.uiuc.edu>
2. <http://gate.ac.uk>
3. <http://web.media.mit.edu/~hugo/montylingua>
2. <http://gate.ac.uk>
2. <http://gate.ac.uk>
5. <http://www.w3.org/RDF>
6. <http://kowari.org>
6. <http://kowari.org>
6. <http://kowari.org>
7. <http://prefuse.org>
8. <http://www.noraproject.org>
9. <http://etext.virginia.edu/eaf/overview.html>
10. Tom Horton, Kristen Taylor, Bei Yu and Xin Xiang, "Quite Right, Dear and Interesting": Seeking the Sentimental in Nineteenth Century American Fiction"*Digital Humanities 2006*, pp. 81-82

document for a presentation, and from that document create slides, notes intended for the lecturer, and notes intended to be part of a hand-out for participants. The system is well documented, with an intended audience being a person who is familiar with TEI and is generally technologically savvy, but may not have much direct experience with complex transformations and stylesheets. Many such TEI users use the commercial XML editor oXygen, which includes most of the needed software packages (e.g., xmllint and Saxon 8) built-in. Therefore instructions are provided for using the system from within oXygen, as well as from the command-line.

The system comprises (in order of academic interest):

- An extensive TEI customization which provides a TEI markup language for presentations, and documentation for the use of that language.
- An XSLT stylesheet that transforms a document conforming to the markup language described by the ODD to XHTML.
- CSS stylesheets for viewing both the source and the various outputs in a browser.
- A shell script front-end (for GNU/Linux, Mac OS X, or similar systems) for executing the commands necessary to generate the system documentation and schemas from the ODD file. This relies on roma.sh, the ODD processing tool written for the TEI by Sebastian Rahtz.[4]
- A shell script front-end (for GNU/Linux, Mac OS X, or similar systems) for executing the commands necessary to apply the XSLT transformations to a source yaps file, generating multiple linked XHTML files.

TEI customization

The TEI customization or ODD file that defines both the yaps schema and its documentation is large (currently almost 2000 lines) and thorough. It demonstrates several interesting features (e.g., deletion of classes, use of Schematron). But more controversially, the documentation therein is not only for the *language* described, but is also the documentation for its *use*. This is certainly not a "pure" or intended use of TEI ODD, and some might even say it is inappropriate. On the other hand, ODD provides a reasonable place for such documentation, and TEI provides a reasonably rich and familiar language with which to write such documentation.

The markup language described by the ODD is much smaller and far more structured than the TEI most of us are used to. From the beginning the goal has been to create a markup language that would be easy to use (because the authors using it are already well versed in TEI), but constructed specifically for the production of slides and associated materials to be used during a presentation. The language is deliberately highly constrained. For example, even in cases where order would be irrelevant as far as processing is concerned, a particular order is required by the schema primarily for the benefit of the author, who has to make fewer guesses about what should come next, and make few if any decisions between equivalent encodings.

The yaps language was not designed as a standalone language and then retrofitted to TEI, but neither were significant compromises to the desired language made to make it easy to express it as a TEI language. Thus the customization ODD is somewhat complicated in order that the end result be a simple, applicable language. For example, the `<body>` element with its `<div>`s is replaced by a `<presentation>` element which has nothing but `<section>` (or `<sectionGrp>`) children. It would have been easier to design a customization which used `<div type="section">`. While this would have been semantically equivalent, it is not possible (in ODD) to provide the desired syntactic constraints on `<div>`, whereas it is easy to provide such constraints on `<section>`.

There are only three possible children of a `<section>`: one for a slide, one for instructor's notes, and one for an accompanying handout. There are currently only 96 elements defined in this language, many of which are metadata elements used in the `<teiHeader>`. As an example of its simplicity, there are only 14 elements available to an author writing the content of a `<slide>`; there are well over twice that many available to the author of a TEI Lite document using `<div type="slide">`.

XSLT

The stylesheet is a single parameterized XSLT 2.0 stylesheet that is over 1000 lines long. It creates relatively abstract valid XHTML output (i.e., most of the input YAPS elements are represented by XHTML `<span>` elements that make use of the `class=` attribute). I am hopeful that it is of high enough quality to be a useful example for beginners; simultaneously I am hoping XSLT experts will be able to provide pointers and advice on improvements.

Further Features

Although the system is quite usable in its current state (Julia Flanders, Christian Wittern, and I have all made use of it for workshops) there are, of course, a lot of improvements still to be made. Many of these improvements I expect to have in place before the conference.

- Make locations of stylesheets, etc., more generalized
- Make licensing conditions (e.g., GPL) more explicit
- Update documentation to match current release of oXygen

- Consider use of a configuration file that would make customization of some of the CSS features easy
- Consider creation of a transform to/from APXL (Apple's schema for Keynote)
- Generate PDF as well as XHTML (not expecting to do this before conference)

1. See <http://www.tei-c.org/>, in particular <http://www.tei-c.org/Consortium>.
2. As of this writing the current alpha release (0.6) is available both via the TEI website and on Sourceforge. See <http://sourceforge.net/project/showfiles.php?group_id=106328&package_id=141127> or <http://www.tei-c.org/P5/>.
3. For the definition of the ODD system, see Chapter 27 "Documentation Elements," *Guidelines for Electronic Text Encoding and Interchange*, ed. Syd Bauman, Lou Burnard, & C. M. Sperberg-McQueen (Text Encoding Initiative Consortium, 2006). <http://www.tei-c.org/release/doc/tei-p5-doc/html/TD.html>. For a tutorial on the ODD system, see Rahtz & Burnard, "One Document Does it all", presented at the TEI annual Members' Meeting 2005, Sofia, Bulgaria. <http://www.tei-c.org/Talks/2005/Sofia/odds.pdf>. For an overview and discussion of the theoretical implications, see Bauman & Flanders, "Odd Customizations", presented at Extreme Markup Languages 2004, Montreal, Canada. <http://www.idealliance.org/papers/extreme/proceedings/html/2004/Bauman01/EML2004Bauman01.html>. For a variety of samples of ODD in use, see Bauman et al., "An Odd Basket of ODDs", presented at DH 2006, Paris, France.
4. Available on Sourceforge at <http://sourceforge.net/project/showfiles.php?group_id=106328&package_id=141128>.

Digital Humanities and the Solitary Scholar

David J. Birnbaum *(djbpitt@pitt.edu)*
University of Pittsburgh
Michael L. Norton *(nortonml@jmu.edu)*
James Madison University
Linda E. Patrik *(patrikl@union.edu)*
Union College
Dorothy Carr Porter *(dporter@uky.edu)*
University of Kentucky
Geoffrey Rockwell *(georock@mcmaster.ca)*
McMaster University
Helen Aguera *(haguera@neh.gov)*
National Endowment for the Humanities

> "The process of creating new knowledge -- of research and scholarship -- is also evolving rapidly away from the solitary scholar to teams of scholars, perhaps spread over a number of disciplines."
> (James J. Duderstadt, "Transforming the University to Serve the Digital Age," CAUSE/EFFECT 20.4 (Winter 1997-98): 21-32.)

In the epigraph for our session, Duderstadt succinctly describes a major challenge facing the humanities – specifically the scholars who make the practice of humanities research their own – as they head into a digital future. The typical scholar in the humanities is generally perceived as a loner, an individual working more or less in isolation towards publications that will be his or her sole responsibility. Scholars are not completely isolated, of course. The sharing of pre-publication knowledge is common, and workshops and conferences abound in all fields to give scholars the opportunity to find out what others are doing, to present their own work to the community, and to talk with others whose interests and research overlap their own. However, as a glance at the contents pages of the most recent issue of any historical or literary journal will suggest, it is not common for more than one humanist to claim authorship for a single piece of scholarship. There is contrast here with scholarship in the hard sciences and engineering, where much work is done by groups, collaboration is the norm, and where several people usually claim authorship to any one publication. Digital humanities inherits from both

of these traditions, encouraging cross-disciplinary collaboration within a community that has traditionally valued independence.

This short paper session will present the experiences of three scholars working more or less in isolation within the context of this collaborative environment. These individuals are alike in that they are all actively involved in the development of digital projects, but their experiences vary widely with regard to time available for working on these projects, their own technical/programming skills and like support available from their institutions, and the availability of funding to support their projects. In addition to these three scholars, we also welcome two others whose work exists to support the efforts of solitary scholars. However, it may be seen that even "solitary" scholars are not so alone in the digital humanities.

"I'm ready to start a project! What do I need to know?"

Our first two speakers represent the technically skilled scholar.

Birnbaum has developed two XML-related projects on his own and he will report on the logistics of these projects. Funding and support for both projects is light and he is responsible for all the technology, although his colleagues provide philological knowledge.

Norton will discuss his experience as a computer scientist with a PhD in Music History developing a project on digital representations of medieval liturgical manuscripts. He is working on the project independently, and while he has the programming and design skills necessary for success, he is facing limitations of time (as he is also trying to write a book, a familiar problem for many).

"Well, how do I learn?"

Our third speaker will discuss alternatives for solitary scholars to learn how to take steps towards developing digital projects.

Rockwell will talk first about how to take advantage of university resources. Many universities have Teaching and Learning Centers and Digital Library Services that you can draw on for help. Some universities have special Humanities Computing units that specialize in supporting digital projects in the humanities. Rockwell will summarize some of the services you can look for at your university.

In addition, Rockwell will speak as a the project leader of the Text Analysis Portal for Research (TAPoR) project (<http://www.tapor.ca/>), which makes text analysis tools available to scholars for integration into their electronic text projects. Rockwell will discuss the TAPoR tools, and will discuss how scholars working in isolation can take advantage of them without the need for programming experience or access to programming support.

"I have the skills, I have a great idea... what about funding?"

Our final two speakers will present two different aspects of acquiring funding for digital humanities projects.

Patrik will discuss issues that she has encountered while attempting to organize a TEI encoding project at a liberal arts college. Although Patrik's institution is supportive of her digital endeavors, she has not had success applying for funds from the major funding agencies. Unfortunately, without access to graduate student labor and a dedicated support staff and with the lack of time that comes from being a full-time professor in an institution where the focus is on undergraduate teaching, Patrik is at a disadvantage. She is interested in exploring the possibilities of collaborating with larger institutions, moving from isolation to collaboration.

Winter will discuss the NEH Digital Humanities Initiative, and how the initiative can support the work of solitary or independent project directors. He will focus on two of the new DH-specific grant initiatives: Digital Humanities Start-Up Grants (<http://www.neh.gov/grants/guidelines/digitalhumanitiesstartup.html>) and Digital Humanities Fellowships (<http://www.neh.gov/grants/guidelines/dhfellowships.html>).

Each of our speakers will have seven to ten minutes to discuss his or her work, making reference to limitations of time, technical support, and funding, and how those limits impact their projects. The remainder of the session will be a roundtable discussion, and the audience will be encouraged to ask questions of the panel and to discuss their own experiences working in isolation

The Paradise Lost Flash Audiotext

Olin Robert Bjork *(olin.bjork@gmail.com)*
University of Texas at Austin
John Peter Rumrich *(rumrich@mail.utexas.edu)*
University of Texas at Austin

In many Digital Humanities projects, the focus has been on acquiring, digitizing, and encoding materials for online search and retrieval, rather than on interface design. A common notion in Humanities computing, carried over from the information technology industry, is that content should be prepared so as to as to be deliverable through pre-existing interfaces as well as those yet to be conceived. But this procedure suits developers more than users. Simply making a wealth of primary and supplemental materials available in digital form does not make them usable. Moreover, the practice of coding and representing humanities materials as information objects flies in the face of many humanities scholars and teachers, who view works of art and culture as historically and materially situated, as unique combinations of form and content. Digital humanists, rather than merely archiving these materials in standard formats, should design interfaces that bespeak their unique qualities and accommodate the desiderata of scholars, teachers, and students.

In designing the interface for our own project, we began with the premise that new media can be used to enhance student comprehension of complex poems, even those for which print resources are rich and abundant. The complex poem we had in mind was John Milton's Paradise Lost (hereafter, PL), perhaps the single literary work most often assigned to English-speaking university students around the world. Whether students run through excerpts from PL in a sophomore survey or pore over the entire epic in an upper-division course, they famously find Milton's poetry difficult to follow. Instructors usually assume that this difficulty owes to its unfamiliar ideas and Milton's intimidating erudition—and surely these are part of the problem. But we have found that when students hear an instructor declaim passages from PL as they follow along in their textbooks, the thrust of the lines suddenly becomes plainer. Recent research on multimedia learning indicates that distinct, additive cognitive pathways mediate the aural and visual reception of language (Mayer 2001). Reading and listening to the same text demonstrably improves understanding and recall.

The aural register is especially crucial for the study of PL. Blind Milton composed his epic orally, dictating it to amanuenses and even insisting that he was relaying what the Muse had first dictated to him as he slept. The original text of PL, then, was not a mute manuscript but a narrative voice attentively heard and transcribed. Even when present-day students successfully comprehend Milton's poetry during a solitary, silent reading, reading through the eyes alone diminishes the aural impact of the verse and the voice as its medium. Instructors and students now seem to be recognizing that the merely visual reading of PL leaves something to be desired. In recent years an increasing number of marathon group readings of Milton's epic has occurred at colleges and even high schools in the U.S.A. and elsewhere. Such voluntary meetings are unlikely to become standard practice any time soon, however, and classroom time does not permit instructors to read aloud or play a recording of the poem (10,565 lines), even if they recognize that comprehension increases markedly when students both see and hear the text.

In the fall of 2004, having thought through this pedagogical situation, we developed an interface design that would correlate an electronic text with an audio track of PL. We then solicited volunteers to record Book Nine, the selection most often anthologized and assigned to undergraduates. In the spring of 2005, with the aid of a Liberal Arts Instructional Technology Services (LAITS) grant, we developed a prototype "audiotext" of Book Nine. This prototype, which is now publicly available at <http://www.laits.utexas.edu/miltonpl/>, uses Adobe/Macromedia Flash technology to synchronize a modernized text with audio and annotations within a single window designed to look like an open book. It uses a karaoke-style moving highlight to indicate the line being voiced and an optional stationary highlight to mark annotated words and phrases. Because Flash is not free software, some may object to its widespread use in digital humanities projects. But developing the audiotext with an open standard such as SMIL and/or SVG would have been far more difficult, if not impossible. Furthermore, Adobe has released the Flash file format specification (SWF), spawning a large and robust open source Flash community as well as third party software that generates SWF files. The SWF format now rivals PDF and HTML in readability and accessibility, respectively. Making the audiotext accessible to users who require screen readers seemed particularly important to us, given a work composed by Milton.

During the Fall of 2005, more than 200 students tested the prototype in various courses: a freshman honors seminar in world literature, a large lecture section of the mandatory sophomore literature survey, and an upper division course devoted to Milton. According to the results of an anonymous survey administered by the instructors, the majority of students (91%) said that they understood the action better than they did

when using a print version only. Such improvement in comprehension implies that students also better understood Milton's notoriously intricate syntax, line by line, and thus became more sophisticated readers overall. These student surveys, as well as comments we received from instructors, testify to the pedagogical utility of our design. But we also believe that our interface provides a better solution for scholars and general readers. Although we have deliberately adhered to the logic and aesthetics of the printed page, our design also uses the capacities of contemporary information technology to improve upon the book. During audio playback, the pages turn automatically, minimizing the cognitive break of page-turning and further encouraging users to continue uninterrupted with the reading.

More importantly, our design solves a problem of print technology that has persisted in digital technology: the tendency of supplemental materials, such as explanatory or textual notes, to disrupt the sort of immersive reading experience vital to the understanding of challenging literary texts. Although notes can enrich a reader's appreciation, when they are printed on the same page as the primary text (footnotes), they can be visually and mentally distracting. On the other hand, if they are printed in an appendix (endnotes), they are difficult to compare with the primary text and are in practice often ignored altogether. Unfortunately, the majority of hypertexts and ebooks have simply reproduced footnote or endnote schemes. Our interface design exploits Flash technology to smooth out this age-old snag in reading, allowing the user to make notes and other supplements appear and disappear as needed, without scrolling, turning pages, or negotiating multiple windows or frames. When the audiotext is in annotation mode, explanatory notes simply replace the text on the page facing the one being read/heard.

In the spring of 2005, we responded to student and instructor feedback by adding two more viewing modes to the prototype. The "comparison" mode displays our modernized reading text in parallel with a diplomatic transcript of the 1674 text on which it is based. This mode allows teachers to point out differences in orthography and punctuation without asking their students to consult a facsimile edition. The "your notes" mode, meanwhile, allows students to edit and save their own notes for each page. These notes are stored on the student's computer in a separate file, from which they are then imported for each new session. This mode thus reproduces a capacity of print books seldom reproduced in electronic formats—space for readers' own annotative marginalia. Teachers who ask their students to keep reading journals may find this mode especially useful. Unlike similar features in Blackboard and other online classroom systems, the notes are individually rather than collaboratively composed, and the mode does not require usernames and passwords. In fact, we encourage users to download the audiotext for offline use rather than load it online in their browsers. The audiotext in this regard represents a departure from the "Web 2.0" movement, a convergence of applications that make the network, rather than the user's own computer, the platform. Although our interface design is applicable to many of these technologies, the audiotext itself provides an example of a richer and more flexible desktop solution for teachers and students than print books or ebooks.

Ultimately, we hope to complete audiotexts of all twelve books. In 2006-2007, with the help of a second LAITS grant, we are developing Books One and Two. John Rumrich has also applied for an NEH Digital Humanities Fellowship that would enable him to complete additional audiotexts in 2007-2008. In the next phase of the project, we plan to develop a separate front-end utility that will search the text of the poem in an XML document that conforms to the P5 guidelines of the Text Encoding Initiative. This utility, which will also be able to search the annotations and user notes, will return results that correspond and connect to specific pages in the audiotexts.

Bibliography

Mayer, Richard E. *Multimedia Learning.* Cambridge: Cambridge University Press, 2001.

The Encoding of Terminology Related to the Medieval Slavic Manuscripts: Philological and Technological Results and Perspectives

Andrej Todorov Bojadžiev
(andreib@slav.uni-sofia.bg)
University of Sofia

The author of the presentation makes an attempt to summarize problems and prospects concerning terminology used in computer supported description of Slavic manuscripts in two Slavic languages – Bulgarian and Russian, and, at the same time – in English and German. The report is a result of a working team in the Bulgarian Academy of Sciences and Sofia University: Anisava Miltenova (Institute of Literature, director of the project), Andrej Bojadžiev (University of Sofia), Margaret Dimitrova (University of Sofia), Irina Kuzidova (Institute of Literature), Regina Koycheva (Institute of Literature), Maya Petrova (Institute of Literature), and Svetla Koeva (Institute of Bulgarian Language). They are working in the frames of the project Metadata and Electronic Catalogues (2004–08), a component of the Repertorium of Old Bulgarian Literature and Letters. The project is oriented to create electronic catalogues and authority files that will serve as integrated repositories of terminological information that has been developed and applied successfully in already existing projects in the realm of medieval Slavic languages, literatures, and cultures. One innovative feature of this project is that beyond serving as a central repository for such information, it will expand the organizational framework to support a multilingual superstructure along the lines of I18N initiatives elsewhere in the world of electronic text technology in general and humanities computing in particular.

The project presumes the possibility of linking the standardized terminological apparatus for description, study, edition, and translation of medieval texts, on the one hand, to authoritative lists of key-words and terms used in bibliographic descriptions, on the other. This will allow the integration of scholarly meta-data and bibliographic references under a single unified framework. Another aim of the project is to create a mechanism for allowing the extraction of different types of indices based upon the imported documents even when the languages of encoding may vary. The primary manuscript description texts are encoded in a TEI-based XML format in the context of the broader Repertorium initiative, and their utility for the type of multilingual authority files, bibliographic databases, and other broad reference resources illustrates the multipurposing that is characteristic of XML documents in the humanities, but on a broader scale than is usual.

One of the project's main aims is to propose an approach with the help of which the main terms of the Medieval culture could be further explained with the original texts, translations and current research activities. In this respect we could distinguish several areas of knowledge: name of the various texts and principles of such a naming, the author of the text(s) and problems related with the transliteration and transcriptions of the names; terms related to the text history its structure and function; the language(s) of the text; the relationship between the text and the codex; the Medieval mankind represented in the texts, etc. This project is based on distinguishing the meanings of particular terms and notions that appear in the text of medieval written documents both within the primary corpus and in comparison to established scholarly terminology.

In the frame of the project several types of indices have been created: names of the texts (in four languages), genre terminology (in the field of hymnography, hagiography, homiletic literature, different kinds of instructions, etc.), types of manuscripts (concerning their function), palaeographic and codicological terms, linguistic terminology These indices could be divided into two types:

- Entries in the form of lists which contain information for usage of particular notion in the original Medieval texts .
- Entries in the form of terminological XML tool, which contain research metadata, in the form of thesaurus.

The main difference between the both is the absence (in the former) and the presence (in the latter) of definitions, and a strict hierarchy of concepts. The first type of the information represents a simple multilingual list of terms. The function of this list could be explained as "It could be the standard way to say this or to write this". The entries of the second type have more sophisticated structure, very close resembling to the encyclopaedic article. A compulsory element of this type of entries is the reference to the authority source, from which the material is extracted and which is very appreciated among scholars. In the informational area the specifications to the particular standards are used.

The project is very actual in the context of discussions in the frames of ISO activities for definitions of markup of the terminological information (cf. TMF: Terminological Markup Framework). Our model is based on the approved standard ISO 12000. The report includes demonstration of the current project's results both in the philological and technical aspects.

Making a Contribution: Modularity, Integration and Collaboration Between Tools in *Pliny*

John Bradley *(john.bradley@kcl.ac.uk)*

Centre for Computing in the Humanities
King's College London

Computing tools have been an issue since the foundation of Humanities Computing, and building modular tools that work together has been recognised as important since at least the CETH meetings in 1995. Geoffrey Rockwell and I first raised issues of modularity in our paper given at the Canadian Learned Societies Conference in June 1992 entitled "Towards new Research Tools in Computer-Assisted Text Analysis". Our proposed tool framework combined data-flow tool modularity with the ability to create pages that mixed scholarly writing with interactive elements like in a *Mathematica* "Notebook". Of course the WWW was sweeping all this away by 1995 and when we took a paper (Bradley, Rockwell 1995) to the CETH "tools" meeting we were merging our modular view with the WWW as it was then emerging.

In our community "modularity" often comes down to the sharing of file formats. Separate tools that can all read the same file format can all operate on common data, and therefore can contribute their particular facilities to the task at hand. File- or Unix-style modularity is a part of what I called the "transformation model" of computing (Bradley 2005), and provides a powerful approach to manipulating data. TuStep (described in Ott 2000) is a splendid example. The model holds a strong attraction within the Digital Humanities community – even extraordinarily creative projects as TAPoR's text analysis portal (TAPoR 2006) or the Nora project (Nora 2006) are, in fact, primarily based on this. However, thinking of modularity primarily in those terms limits our views to those of computing about 25 years ago – before the advent of the graphical user interface (GUI).

The GUI radically changed the way we think about computing, and even today, more than 25 years later, its significance continues to reverberate. It is surely true to say that most humanists have the GUI as a model driving how they think of their computer. The impact has been twofold:

- First, users now expect that the computer screen will allow them to directly manipulate materials that interest them. Think of the words in a word processor.
- Second, interaction between tools needs to be possible not only between files, but also on the screen as well. Users expect to be able to incorporate parts of spreadsheets inside word documents, for example.

To reflect this broader sense of collaboration between tools, we need some different language. In this paper I will use the word *integration* for that aspect of tool collaboration that focuses on GUI issues. The use of *integration* might make us also think of the integrative nature of humanities scholarship – an intentional parallel.

Integration in the GUI presents challenges. First, the development of tools allowing for direct manipulation of objects on the screen is more complex and costly than resolving file-sharing issues. Furthermore, if independent tools are to interact on the screen – elements maintained by potentially different tools sharing screen space – then they must operate in a computing framework that makes this kind of thing practical.

Pliny (Pliny 2006) has been developed precisely to draw attention to these two issues and to encourage some thinking about cooperating tools beyond file-oriented "modularity". It both significantly broadens how computers can support humanities research and suggests a much richer and more acceptable interface to those tools that might well attract a larger number of scholars. *Pliny* tackles this in two ways. First, it is supports annotation and note taking – functions central to several aspects of humanities scholarship and which I believe have been largely neglected by the DH community – and, second, it is built using a framework called Eclipse (Eclipse 2006) which is deliberatively created to support GUI-level interaction between tools of the kind I mentioned above. Eclipse also supports rich collaboration between tools developed by independent developers.

Personal annotation and note-taking compels us to think about tool integration because it is widespread in humanities scholarship and runs across all kinds of scholarly work. Scholars write notes to record their reactions to not only books they read, but also web pages they view. Furthermore, if they have tools that do (say) text analysis, they would probably want to record notes about that as well. The provision of note-taking within a *particular* website is insufficient for personal note-taking (although it might well fulfil a useful need supporting public comment about the material the website offers) because most scholars work across a broad range of materials, and their note taking capability must reflect this. Notes from a book will at some point need to be brought in contact with notes about, say, an online archive or a journal, or about findings from the text

analysis tool. Personal scholarly note taking integrates by its very nature.

Personal annotation also draws our attention back to the *software application* as the context in which tools can and should be built, rather than thinking of the browser as the context for tool delivery. It is surprising how difficult it is to make this point to those in the Digital Humanities. Of course part of the reason is that XML, one of the key technologies that fuels some portion of the DH community, has been developed specifically to work in the context of the WWW. However, if we wish to develop a more broadly based humanities community who use technology in more sophisticated ways to support their research we need to broaden our focus beyond the WWW. Providing scholarly resources on the WWW supports scholars, of course, and there remains some excitement that they can get materials readily right to their desktop. However, as I argued in Bradley 2005, once on that desktop all they can do with them is read them on the screen or print them out – a webpage, even one designed with all the clever parts of CSS and AJAX, allows only manipulation within its own page context. It contributes little to the integrative aspect of personal scholarship which must, by its very nature, often bring materials together from disparate sources.

Software Applications work with personal materials. Surely this is a central element of humanities scholarship. A word processor creates materials that *belong* to the researcher. Until web browsers can be used to create materials that are stored as personal data – on the user's own machine – they cannot replace applications (see a similar assertion from a technical perspective in Charland 2005). Even the XML folk – part of the community driving a WWW view of humanities computing because of XML's compatibility with website creation – use an application such as Oxygen to create XML materials in the first place. For the same reason that the creation of an XML file involves an editor rather than a web browser, the creation of personal notes – especially those gathered from reactions across a great range of sources – requires a personal application. Furthermore, more than one study into computing and humanities scholarship (see both Brockman *et al* and Siemens *et al* 2004) has found that scholars have tried to apply inappropriate applications such as word processors to support this need, and have not been very satisfied with the result.

Applications can be built in several different frameworks. *Pliny* is written in Java using the Eclipse framework rather than the one provided by Sun to create desktop tools. I didn't choose Eclipse because it was easy for me to use it – indeed I had to learn it from the beginning during the past year or so. I choose Eclipse because it is designed specifically to support the integration of tools developed by separate developers. Eclipse provides a way to share out screen space between windows managed by different tools. Furthermore, it also provides mechanisms for elements from separate tools to share the same window (called "making a contribution" in Eclipse). Tools built with the Eclipse plug-in model need not operate in isolation from other related tools. A text analysis tool built in Eclipse could use *Pliny* elements to allow users to record notes while using the TA tool, and the notes that were recorded in this way would also be visible in the context of notes created with other materials – say from reading a book, or annotating a web page.

So, in this presentation I have an impossible task. First, I am promoting the idea that we focus more on application development than web development, a technically more demanding activity, and one that goes against the grain of much work in DH over the last decade. Second, for those who might subscribe to this idea, I am promoting that we build our tools according to the Eclipse model rather than the more widely understood Sun/Java framework. Perhaps I won't convince anyone here. However, I believe that unless we start to think of tools in the context of applications, as our scholarly community does, and unless we start to think more seriously of tool building in the context of GUI integration in addition to data sharing, we will never get the attention of most scholars in the humanities.

Bibliography

Bradley, John. "What You (Fore)see is What You Get: Thinking about Usage Paradigms for Computer Assisted Text Analysis." *Text Techology* 14.2 (2005): 1-19. Accessed 2006-09-01. <http://texttechnology.mcmaster.ca/pdf/vol14_2/bradley14-2.pdf>

Bradley, John, and Geoffrey Rockwell. "Towards New Research Tools in Computer-Assisted Text Analysis." Paper presented at The Canadian Learned Societies Conference, June 1992. 1992. <http://www.cch.kcl.ac.uk/legacy/staff/jdb/papers/learneds.html>

Bradley, John, and Geoffrey Rockwell. "The Components of a System for Computer Assisted Text Analysis." Paper presented at the CETH Workshop on Future Text Analysis Tools, October 1995. 1995. <http://www.cch.kcl.ac.uk/legacy/staff/jdb/papers/ceth95.html>

Brockman, William S., Laura Neumann, Carole L. Palmer, and Tonyia J. Tidline. *Scholarly Work in the Humanities and the Evolving Information Environment.* Washington, D.C.: Digital Libary Federation and Council on Library and Information Resources, 2001.

Charland, Andre. "Will Ajax Replace the Desktop?" *developer.com.* 2005. Accessed 2006-10-01. <http://www

.developer.com/design/article.php/10925_3574116_1>

Eclipse. 2006. Accessed 2006-10-01. <http://www.eclipse.org/>

Nora. 2006. Accessed 2006-10-01. <http://www.noraproject.org/>

Ott, Wilhelm. "Strategies and Tools for Textual Scholarship: the Tübingen System of Text Processing Programs (TUSTEP)." *Literary & Linguistic Computing* 15.1 (2000): 93-108.

Pliny. 2006. Accessed 2006-10-01. <http://pliny.cch.kcl.ac.uk/>

Siemens, Ray, Elaine Toms, Stéfan Sinclair, Geoffrey Rockwell, and Lynne Siemens. "The Humanities Scholar in the Twenty-first Century: How Research is Done and What Support is Needed." *ALLC/ACH 2004 Conference Abstracts.* Göteborg: Göteborg University, 2004.

TAPoR: Text Analysis Portal for Research. 2006. <http://test-tapor.mcmaster.ca/portal/portal>

Spatially Enabling RiverWeb, a Web-Based Resource for Historical Exploration of the American Bottom

Vernon Burton *(vburton@ncsa.uiuc.edu)*
NCSA
University of Illinois at Urbana-Champaign
Luc Anselin *(anselin@uiuc.edu)*
NCSA
University of Illinois at Urbana-Champaign
Simon Appleford *(sapplefo@uiuc.edu)*
NCSA
University of Illinois at Urbana-Champaign
Myunghwa Hwang *(mhwang4@uiuc.edu)*
Spacial Analysis Laboratory
University of Illinois at Urbana-Champaign
James Onderdonk *(onderdon@uiuc.edu)*
NCSA
University of Illinois Urbana-Champaign

RiverWeb (<http://www.riverweb.uiuc.edu>) is a web-based education and outreach program to promote environmental education and historical awareness about rivers and their watersheds. One aspect of the project is an extensive collection of historical materials pertaining to early settlements in the “American Bottom,” a location just south of the confluence of the Mississippi, Illinois and Missouri rivers, near modern-day St. Louis. This also includes a large historical record of the early development of East St. Louis (IL), from its early settlement through the twentieth century, up to and including the most recent census.

This poster illustrates efforts to spatially enable the RiverWeb collection. The outcome is a dynamic web mapping application that forms a browser-based flexible user interface to the collection of historical materials, newspaper clippings, directories and census information. The system has been built using open source software compatible with Open GIS Consortium (OGC) standards and includes web mapping functionality, a gazetteer and basic geovisualization. The paper describes the design and architecture of the system and the

implementation of the linkage between the historical information and its spatial imprint. This is illustrated with a historical analysis of neighborhood change in East St. Louis as a result of redlining policies in the early 20th century. Change in neighborhood profile over time and across space can be visualized and basic exploratory spatial data analysis (ESDA) techniques provide ways to quantify these patterns.

Distributed Multivalent Encoding

Paul Caton *(Paul_Caton@brown.edu)*
Brown University

Distributed multivalent encoding (DME) describes a web-based text encoding practice and the result of that practice. It assumes a digital text resource with a browser interface that allow users to associate N encodings with N texts in the resource. Users need have no other connection with the resource than their using it and constraints should be minimal; hence the encoding is *distributed* among multiple creators and multiple conceptual approaches. To realize their encoding(s) users may ignore existing encodings; they may apply one tagset to one text, or to multiple texts, or they may apply multiple tagsets to one text, or multiple texts; any element they create can reference any other element created by themselves or by someone else. Thus the resource's encoding becomes *multivalent* in part and in whole.

A cluster of activities defines the space of distributed multivalent encoding, including annotation, keywording (including "folksonomizing"), editing, and humanities criticism. For any one component of DME anticipatory previous work exists.[1] Any newness of DME lies only in the surprising fact that the components have never been fully assembled, though they have all been present for some years. Now, however, we can see DME emerging, albeit in hesitant, not-fully-developed forms.[2]

A DME resource has a base text (or texts), a web interface for creating encoding, and a mechanism for storing encoding; these three things enable all subsequent activities: editing, deleting, retrieving, searching, etc. Considering each of the three parts in a little more detail we will see few real technical obstacles in DME's path.

Distributed multivalent encoding starts with a reference base digital text. A reference base text should be clearly marked as such by the resource managers. Only they may change it (or give someone else permission to change it), and any change must be well publicized. A reference base text must have a well-defined beginning and end, and a well-defined internal base reference structure. While it is certainly possible to regard either the simple byte sequence or the character sequence as the base reference structure and then link encodings to byte or character offsets, this is not a robust solution. Practicality suggests the internal base reference structure should itself be an encoding; convenience and practicality further suggest a

presentation-based encoding, on the grounds that presentation in a medium reflects a community's sense of that medium's main communicative organization and delivery units (Caton, 2001). A TEI Lite encoding (<http://www.tei-c.org/Guidelines2/index.xml.ID=lite>), for example, or one conforming to DLF Level 4 (<http://www.diglib.org/standards/tei.htm#level4>), would be suitable. We should stress, though, that the reference structure makes no claim to being representationally definitive. Within the structure, distinctions between tags and #PCDATA are *purely structural* and do not define either "text" (as general phenomenon) or "the text". A supplementary encoding – ie. an encoding created by a user – associates with *the text as represented within the reference structure*, not "the text" as some reified cultural object. If we follow the reasoning of Renear et al on the relationship between FRBR entities and XML documents, we would probably consider the base reference text a *manifestation* of an expression, especially because we incline towards a presentation-based encoding (which, as Caton argues, is what OHCO-style encoding *is*)(Renear et al, 2004, Caton 2001). The FRBR vocabulary, however, hardly resolves the problematic semantics of the common phrases "the text" and "a text". Hence our insistence on treating the internal reference structure as definitive only within the DME resource as a system. Indeed the very point of a DME resource is to allow users to create encodings that they can treat as definitive *for their purposes*.

The web interface presents the base text to the user and allows the user to associate encoding(s) with parts of the text. The encoding must allow for multiple overlapping hierarchies, and so a format such as CLIX/TEI HORSE should be used (DeRose, 2004, Bauman, 2005). While the interface programming might be complex in terms of having to manage numerous details, the required functionality is straightforward. The actual mechanics a resource employs are not important, except as regards the degree to which they make the process awkward. They will vary according to programmers' preferences and with changes in technology. The proof-of-concept Limner interface, for example, uses HORSE and relies on explicit element IDs and user-selected strings to mark where start and end tags go. DeRose notes that '[o]ne often hears that … IDs are somehow "safe" pointers into documents (DeRose, 2004). However, this is not true; they are at most "safer" than many other methods.' His point is well taken. However, there must be *some* reference system, and unless we resign ourselves to inline markup and unwieldy file sizes, it seems preferable to keep the supplementary markup separate from the base text and use a system that is (to view the cup as half full rather than half empty) *at least* safer than many other methods. A combination of full XPath plus element ID plus a string of sufficient length to have a strong chance of being unique should allow a DME resource to consistently associate a stored out-of-line element with its proper position with respect to the internal base reference system. The Limner implementation is rather crude and already dated; the DOM scripting features of current browsers together with wider availability of XPath handling functions in programming languages offer many opportunities to improve upon it.

The actual details of storage are also of limited importance. Relational databases can hold the information (as with Limner) but it seems likely that future DME resources will use native XML databases.

Without downplaying the amount of work involved, we can confidently say that DME is perfectly possible with current technology and that DME resources will be built in the near future. The real unknowns (and potential problems) lie on the social side. Who gets to encode? Will all supplementary encodings be equal, or will some be "more equal" than others? Will encodings be moderated? Will differently encoded and competing base reference texts proliferate until the very notion of a base reference becomes utterly compromised? Will a class system of resources emerge, driven by a scholarly fear of non-scholars' contributions? The relative success of Wikipedia in the face of all the things that could have stopped it should make us optimistic. Probably DME will initially develop in constrained forms, with resources authorizing users and retaining ultimate editorial control over supplementary encodings. In time we hope to see distributed multivalent encoding become a widespread, democratic practice

---

1. For example, the resource based around Pico Della Mirandola's Conclusiones CM publicae disputandae (PICO) features an annotation system which is also employed by the Virtual Humanities Lab (VHL), a resource whose development plan includes implementing a form of DME.
2. See LIMNER, for example: a site intended as a proof-of-concept DME resource, still in an early stage.

Bibliography

Bauman, Syd. "TEI HORSEing Around." *Proceedings of Extreme Markup Languages 2005, Montréal, Québec, August 2005.* 2005. <http://www.mulberrytech.com/Extreme/Proceedings/html/2005/Bauman01/EML2005Bauman01.html>

Caton, Paul. "Markup's Current Imbalance." *Markup Languages: Theory and Practice* 3.1 (2001).

DeRose, Steven. "Markup Overlap: A Review and a Horse." *Proceedings of Extreme Markup Languages 2004, Montréal, Québec, August 2004.* 2004. <http://www.mulberryte

ch.com/Extreme/Proceedings/xml/2004/DeRose01/EML2004DeRose01.html>

Digital Library Federation. *TEI Text Encoding in Libraries: Guidelines for Best Encoding Practices.* Version 2.1. 2006. <http://www.diglib.org/standards/tei.htm#level4>

LIMNER . . <http://golf.services.brown.edu/projects/Limner/>

PICO . <http://www.stg.brown.edu/projects/pico/index.php>

Renear, Allen H., Pat Lawton, Christopher Phillippe, and David Dubin. "An XML Document Corresponds to which FRBR Group 1 Entity?" ." *Proceedings of Extreme Markup Languages 2004, Montréal, Québec, August 2004.* 2004. <http://www.mulberrytech.com/Extreme/Proceedings/html/2003/Lawton01/EML2003Lawton01.html>

TEI Lite . <http://www.tei-c.org/Guidelines2/index.xml.ID=lite>

Virtual Humanities Lab (VHL) . . <http://golf.services.brown.edu/projects/VHL/>

The WWW as Curricular Method in the Digital Humanities

Tatjana Chorney *(Tatjana.Chorney@SMU.ca)*
Saint Mary's University

The National Panel Report released by the Association of American Colleges and Universities (2002), calls for a "dramatic reorganization of undergraduate education" to address the challenges faced by higher education in a time of transformation from an industrial to a knowledge-based society (vii). The report states that "education practices invented when education served only the few are increasingly disconnected from the needs of contemporary students"(viii) and the demands of citizens of a diverse and interconnected world.

The Report recommends an invigorated and practical liberal education offering knowledge that all students, regardless of backgrounds, fields, or chosen higher education institutions, should acquire. The college student of the twenty-first century needs to become an "intentional learner" who can thrive in a complex world, and who can adapt to new environments, integrate knowledge from different sources, and transform information into knowledge and knowledge into judgment and action (xi.). The Report urges an "end to the traditional, artificial distinctions between liberal and practical education" and advocates a kind of instruction and learning that looks beyond the classroom to the world's major questions" (xii).

The changing nature of colleges and universities and the reconstruction of education it calls for is in great part conditioned by what Douglas Kellner calls the "Great Transformation," powered by one of the most dramatic technological revolutions in history (2003, 51). The revisioning of higher education is increasingly seen in correlation with the development of digital technologies, which are changing not only traditional models of work, leisure and communication, but also the nature of knowledge itself, as well as models of acquiring and processing information cognitively. In the Report it is asserted that the "intellectual and practical skills that students need are extensive, sophisticated and expanding with the explosion of new technologies" (xi). Like the AACU Report, Kellner too argues that traditional and specialized aspects of education need to be overcome in order to develop alternative pedagogies and multiple "literacies" to meet the challenges of an interconnected global society. The study of interrelationships, connectivity, transfer, and integration, leading to the development of critical judgment, is proposed as the basis of the new liberal curriculum in all disciplines of the humanities.

The five key concepts for the new curriculum are in fact not so much concepts as they are markers of cognitive processes. As such, the reform of higher education is to take place primarily in terms of methodology of teaching. Given that "hypertext is a mental process, as well as a digital tool" (Gilster, 137), and the WWW is "an embodiment of human knowledge"(W3C), exploring the relationships between cognition and technology in the context of the new humanities and pedagogy can be useful. I would like to suggest that the digital media offer a useful concrete, but also a cognitive tool for teaching the five processes that are to be the core of the new humanities. This claim has theoretical and practical implications. Theoretically, it calls for becoming more aware how the computer is altering our ways of engaging with specific disciplinary questions cognitively and methodologically (McCarty, 1). My concerns in this paper, however, have to do with the practical applications of the impact of digital media on cognition within the area of the humanities. Practically, the claim calls for explicit instruction within the context of each discipline in the methods of organization and manipulability that underlie the presentation of material on the WWW.

The main vehicle of the digital media, the WWW, in its nature embodies, illustrates and enables through its functioning all four of the processes suggested as the basis for the new liberal curriculum--interrelationships, connectivity, transfer, and integration. By their very nature "the new media technologies externalize and objectify reasoning" (Manovich, 59). The WWW resists attempts at standard systematization, and demonstrates the co-existence of and interrelationships between multiple and apparently contradictory perspectives on a single issue. One GOOGLE keyword search will retrieve hundreds of documents linked by one single term, but applied variously in different contexts, emphasizing the importance and nature of connectivity, transfer and integration.

In relation to a given subject of inquiry or task, the non-linearity in the presentation of material and ideas on the WWW encourages "intentional" involvement on the part of learners, as there is no longer one "solution" or a single "interpretation," but a variety, all situated within their own context and knowledge. In order to find solutions to given questions or problems, learners have to engage in a process of discovering connections among apparently disparate materials and contexts, then find ways of transferring and integrating parts of materials into a new context. The WWW also offers alternative models of grouping materials, such as scaffolding, and it promotes the idea that conceptual knowledge cannot be separated from the contexts in which it is represented (Wiles and Littlejohn, 2003; Campbell, 2004; Cole, 2000; Carr, 1998;). The ready availability of various information on the WWW turns research and interpretation not so much into an exercise that depends upon finding information, but one that strongly emphasizes cognitive operations depending on critical thinking: classification of it (finding meaningful interrelationships) and making use of it (transferring it) by arranging it meaningfully in a give context (integrating it). Scholarship and instruction in the humanities has always relied on these processes; however, with computer connectivity and the speed with which these processes happen, the WWW amplifies them, enables them "physically" and "on-demand" and thus makes them more explicitly and self-consciously "teachable" than before.

In addition, because the WWW offers multiple presentations of information, it illustrates that knowledge and heuristics are not absolute, but situated within various communities of knowing. Conceptually there is no "closure" or "ending" online, but rather a constant process of evaluation of materials (Rhodes and Sawday, 12) that are open to revisions, additions and remodeling. Becoming aware of the implications the nature of the WWW has for understanding what constitutes knowledge, argument, opinion, analysis and interpretation, leads to the development of critical discernment, evaluative capacity and judgment.

While traditional "mass education tended to see life in a linear fashion based on print models and developed pedagogies which broke experience into discrete moments and behavioral bits," a new critical pedagogy of the digital humanities could produce skills that "enable individuals to better navigate the multiple realm and challenges of contemporary life" (Kellner, 9). Making the WWW, its capabilities and operating functions explicit models of intentional learning can help educators in the digital humanities illustrate how knowledge in the humanities is positively affected by the digital medium, and how the new pedagogy leads to learning as an active, social process bound up with experience connected with wider socio-political paradigms of change.

Bibliography

Greater Expectations: A New Vision for Learning as a Nation Goes to College: National Panel Report. Washington, D.C.: Association of American Colleges and Universities, 2002. <http://www.greaterexpectations.org>

Campbell, Kathy. *E-ffective Writing for E-Learning Environments.* Hershey, PA: Information Science Publishing, 2004.

Carr, Kevin Michael. Dissertation. University of Idaho, 1998.

Cole, Robert A., ed. *Issues in Web-Based Pedagogy: a Critical Primer* . The Greenwood Educators Reference Collection. Westport, CT: Greenwood Press, 2000.

Gilster, Paul. *Digital Literacy.* New York: John Wiley, 1997.

Kellner, Douglas. "Toward a Critical Theory of Education." *Democracy and Nature* 9.1 (2003): 51-44.

Manovich, Lev. *The Language of the New Media.* Cambridge, MA: MIT Press, 2001.

McCarty, Williard. *We Would Know How We Know What We Know: Responding to the Computational Transformation of the Humanities.* Accessed 2004-05-06. <http://digitalhumanities.org/views/Essays/WMcCartyComputationalTransformation>

Rhodes, Neil, and Jonathan Sawday, eds. *The Renaissance Computer: Knowledge Technology in the First Age of Print* . London: Routledge, 2000.

Wiles, Kathy, and Allison Littlejohn. "Supporting Sustainable E-Learning: A UK National Forum." *Interact, Integrate, Impact: Proceedings of the 20th Annual Conference of the Auatralasian Society for Computers in Learning in Tertairy Education.* Ed. G. Crisp, et al. Adelaide, Australia, 2003.

World Wide Web Consortium (W3C). *Definition of WWW*. retrieved from: <http://www.bitpipe.com/tlist/World-Wide-Web.html>

Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project

Arianna Ciula *(arianna.ciula@kcl.ac.uk)*

Centre for Humanities Computing
King's College London

Paul Spence *(paul.spence@kcl.ac.uk)*

Centre for Humanities Computing
King's College London

José Miguel Vieira

Centre for Humanities Computing
King's College London

Gautier Poupeau

École Nationale des Chartes
Université Paris-Sorbonne

This paper will focus on the use of technologies traditionally associated with knowledge management and ontological representation to express complex associations between entities in historical texts that have been marked up in XML according to the Text Encoding Initiative guidelines.[1] In particular, we will describe our exploration of the potential use of RDF (Resource Description Framework)/OWL (Web Ontology Language) technologies and will reflect on the role of an ontology in facilitating the interpretation of implicit and hidden associations in the sources, examining its use and limits in a digital humanities project in connection with editing tools and delivery issues.

We will demonstrate our findings based on the Henry III Fine Rolls project,[2] where an RDF ontology is being developed to make explicit information about person, place and subject entities marked up as “instances” in the core texts themselves.

The Henry III Fine Rolls project and the need for authority lists

The Henry III Fine Rolls is a three-year collaborative project between King's College London and the National

Archives (UK) that aims to represent the complexity of a historical object known as the 'fine rolls' which chart offers of money made to King Henry III of England in exchange for a wide range of concessions and favours. A total of 64 rolls containing around 800 parchment membranes, one for almost all of the fifty-six years of Henry III's reign from 1216-72, survive in the UK National Archives.[3] Each fine roll was compiled in Latin by a handful of scribes and taken as a body of documentary evidence, the rolls are of "prime importance in the study of political, social, and economic history and of government and administration at a local and national level."[4]

The project will cover the first thirty-two years of Henry III's reign down to 1248 and will result in both print and digital editions of the rolls, using as a model the format of the traditional printed 'calendar' (an English summary of records, plus a set of indices) in connection with the digitised images of the rolls themselves. The digital edition will have a sophisticated search interface and both print and digital editions will provide a series of indices for people, places and subjects that include complex associations between the various entries.

The core texts were encoded in XML using the TEI guidelines and include mark-up related to some aspects of the physical structure of the rolls as material artefacts.[5] Particular attention has been given in the mark-up to the occurrences of names of persons, places and institutions. Furthermore, the project researchers have identified and marked-up relevant subjects in the fine rolls. Therefore, while the general mark-up in TEI XML provides a framework for theoretical analysis and practical encoding of the text of the calendar as it is being edited and summarised in English, the need for additional modelling has emerged so as to:

- associate textual instances of the same person, subject occurrence or place to their correspondent logical authority —whether that be a person (identified or anonymous), a subject class or a place;
- express the mutual relationships between the relevant authorities (e.g. individuals, locations, institutions and subjects).

Why RDF/OWL?

In previous work on projects requiring an expression of the associations between sources in core texts marked up using TEI XML we have taken more traditional approaches using XML structures or relational databases to solve the problem, but have increasingly found these wanting. There has been some research in the TEI community in this area lately, particularly around 'biographical and prosopographical data'[6], but at the time of writing this was not mature enough for us to make a commitment to using it and in any case our attention was drawn to other standards whose main objectives are closer to what we are trying to achieve.

After conducting a comparative evaluation of possible standards to create authority lists structures that included research into RDF/OWL, Topic Maps and MADS (Metadata Authority Description Schema), we opted for RDF/OWL.

The main reasons for this choice were the following:[7]

- RDF/OWL models provide a logical organisation of data together with the possibility of a flexible manipulation of meanings (e.g. rich expression of relationships among objects/entities mentioned in the source text, where an object/entity might be a person, a place or a subject);
- RDF/OWL decreases ambiguity by identifying unique entities independently from their instances in a decentralised way.

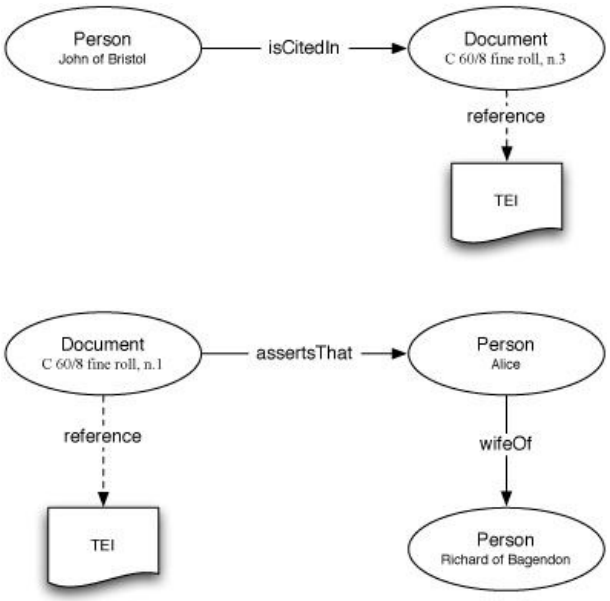
While the more practical advantages are that:

- as a set of standards which are at the heart of the Semantic Web, RDF/OWL is internationally recognised and supported - its models could facilitate the interconnection between humanities computing projects in general and between directly related data in particular;
- there is a relatively good selection of tools for RDF/OWL compared to the other technologies that we have explored;
- RDF/OWL can be expressed in XML format, thus allowing the re-purposing of data for web delivery (e.g. creation of indices through the use of XSLT).

Ontology: Structure, Tool and Delivery

The RDF schema and OWL are used to define the knowledge domain around the core materials on the project (the fine rolls) using classes and predicates. Already existing predicates are used as extensively as possible (e.g. CIDOC-CRM, Dublin Core, Friend of a Friend, Simple Knowledge Organization System).

The connection between the TEI source files and the ontology play an important role in our model.[8] Indeed the XML files not only feed the ontology with some data (e.g. variants of a certain person name), but they are themselves part of the ontological model. The fact that they exhibit names and assert facts related to those names can be modelled in the ontology together with the statement that a person name (proper name or in general reference string) is associated to a specific person as an instance of the entity person (e.g. that has a gender, a possible status etc.). Quoting Eide,[9] we include "explicit statements of the sources of the assertions exhibited in the text":



The concrete connection between the XML files and the ontology is maintained through the use of identifiers, as in the example below:

```
<persName key="arsic_robert"> Robert Arsic </persName>  →  <frh3:Man rdf:ID="arsic_robert"> … </frh3:Man>
```

Evidently, the value and granularity of the ontology depends on various factors, which include the richness of the source files themselves (e.g. how well do the fine rolls express the reality of the roles/occupations in the thirteenth century?), the specific interests of the researchers (e.g. in the types of fines rather than in the amount of money they keep record of) and the definability of the knowledge domain itself (as shown in the contrast between the identification of individuals and the much more blurred identification of subjects).

In addition to expressing complex associations in an abstract intellectual sense, it was also crucial to the project to create a system endowed with sophisticated facilities for editing and final publication. We will describe our experience in creating such tools, in particular taking into account such project requirements as:

- the facilities for managing/editing data: the addition of new information, new classes/predicates/literal values; the definition of relationships between entries, the possibility to merge instances of classes or subclasses (e.g. while editing we may realise that person A and B are actually the same);
- the connection and synchronisation between the ontology structure and data on the one hand, and the TEI source files on the other (e.g. links from classes in authority list to actual references to those classes as exhibited in the mark-up);
- the facility to browse information alphabetically or categorised in some way (e.g. by date, county, role, kind of relation) and to search for particular associations as expressed in the ontology (e.g. all the locations connected to the person called 'Robert Arsic' or all the relatives of Robert Arsic).

Conclusions

In the case of the fine rolls (as is true of many projects involving complex primary sources) it is not enough to mark-up an occurrence of a person name if you wish to create complex indices or to create structured search functions.

Moreover, an authority-based approach is essential in order to make the resource interoperable; external authority lists are needed to record information in a systematic, comprehensive and possibly standardised way, so as to:

- store additional information related to persons, places and subject (exhibited and marked up as corresponding instances in the TEI XML files);
- make explicit the multiple connections between places, persons and subjects;
- merge or disambiguate identifications and eventually correct the original mark-up of the rolls.

Our paper will describe how, for the Henry III Fine Rolls project, an RDF/OWL ontology was used to model complex associations and how this has assisted the project researchers in the interpretation of the material they are editing and facilitated the production of a digital resource that will allow future users to browse the material under different perspectives, to explore the relationships among mentioned individuals, locations and subjects, and to foster new understandings.

---

1. <http://www.tei-c.org>
2. <http://www.frh3.org.uk>
3. <http://www.nationalarchives.gov.uk/>
4. Dryburgh, Paul. “Henry III Fine Rolls Project” (paper presented at the Institute of Historical Research, London , February, 9, 2006).
5. For a more detailed presentation of the mark-up model see Ciula, Arianna. “Searching the Fine Rolls: A Demonstration of the Electronic Version” (paper presented at the International Medieval Congress, University of Leeds, July 10-13, 2006).
6. See “Biographical and Prosopographical Data”. In Sperberg-McQueen, C. M., and Burnard, Lou, eds. "TEI P5 Guidelines for Electronic Text Encoding and Interchange". <http://www.tei-c.org/release/doc/tei-p5-doc/html/ND.html#NDPERS> (accessed 15 November 2006).

7. OWL is a language developed on top of RDF by W3C to write ontologies. For W3C RDF specifications on RDF see <http://www.w3.org/RDF/>. For an overview and further resources on OWL see <http://www.w3.org/2001/sw/>
8. For a similar approach see Eide, Øyvind and Ore, Christen-Emil. “TEI, CIDOC-CRM and a Possible Interface between the Two” (paper presented at Digital Humanities 2006, Université Paris-Sorbonne, July 5-9, 2006) <http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf>.
9. Eide, Øyvind. “The Exhibition Problem. A Real Life Example with a Suggested Solution” (paper presented at Digital Humanities 2006, Université Paris-Sorbonne, July 5-9, 2006) <http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf>.

Bibliography

Boot, Peter. "Decoding Emblem Semantics." *Literary & Linguistic Computing* 21(Supplement 1) (2006): 15-27.

Ciula, Arianna. "Searching the Fine Rolls: A Demonstration of the Electronic Version." Paper presented to the International Medieval Congress 2006, University of Leeds, July 10-13, 2003.

Dryburgh, Paul. "Henry III Fine Rolls Project." Paper presented to the Institute of Historical Research, London , February 9, 2006.

Eide, Øyvind. "The Exhibition Problem. A Real Life Example with a Suggested Solution." Paper presented at Digital Humanities 2006, Université Paris-Sorbonne, July 5-9, 2006. Abstract available at <http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf> (accessed 15 November 2006).

Eide, Øyvind, and Christian-Emil Ore. "TEI, CIDOC-CRM and a Possible Interface between the Two." Paper presented at Digital Humanities 2006, Université Paris-Sorbonne, July 5-9, 2006. Abstract available at <http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf> (accessed 15 November 2006).

Poupeau, Gautier. "De l’index nominum à l’ontologie. Comment mettre en lumière les réseaux sociaux dans les corpus historiques numériques?" Paper presented at Digital Humanities 2006, Université Paris-Sorbonne, July 5-9, 2006. Abstract available at <http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf>(accessed 15 November 2006).

Spence, Paul. "The Henry III Fine Rolls Project." Paper presented at Digital Humanities 2006, Université Paris-Sorbonne, July 5-9, 2006. Abstract available at <http://www.allc-ach2006.colloques.paris-sorbonne.fr/DHs.pdf> (accessed 15 November 2006).

"Biographical and Prosopographical Data." *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Ed. C. M. Sperberg-McQueen and Lou Burnard. Accessed 2006-11-15. <http://www.tei-c.org/release/doc/tei-p5-doc/html/>

Tuohy, Conal. "Using XML Topic Maps to Present TEI." Paper presented at the 5th TEI Meeting, Sofia, Octobery 28-29, 2005.

'Something that is interesting is interesting them': Using Text Mining and Visualizations to Aid Interpreting Repetition in Gertrude Stein's *The Making of Americans*

Tanya Clement *(tclement@umd.edu))*
University of Maryland, College Park
Anthony Don *(don@cs.umd.ed,)*
Human Computer Interaction Lab (HCIL)
University of Maryland, College Park
Catherine Plaisant *(plaisant@cs.umd.edu)*
Human Computer Interaction Lab (HCIL)
University of Maryland, College Park
Loretta Auvil *(lauvil@ncsa.uiuc.edu)*
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign
Greg Pape *(gpape@uiuc.edu)*
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign
Vered Goren *(goren@uiuc.edu)*
National Center for Supercomputing Applications (NCSA)
University of Illinois at Urbana-Champaign

With *The Making of Americans*[1], Stein's goal was to record "being" as it is manifested in repetition. Lauded by some critics who thought Stein accomplished what T.S. Eliot demanded of all writers—to make art, literature, and language "new"—she was also criticized by others like Malcolm Cowley who said Stein's "experiments in grammar" made her novel "one of the hardest books to read from beginning to end that has ever been published."[2] More recent critics have attempted to aid interpretation by charting the correspondence between structures of repetition and the novel's discussion of identity and representation. Yet, the use of repetition in *The Making of Americans* is far more complicated than manual practices or traditional word-analysis programs (such as those that make concordances or measure word-frequency occurrence) could indicate. The text's large size (almost 900 pages and 3183 paragraphs), its particular philosophical trajectory, and its complex patterns of repetition make it a useful case study for analyzing the interplay between the development of text mining tools and the hermeneutics employed in interpreting literary texts in general.

The reason that text mining procedures might be particularly illuminating for Stein's *The Making of Americans* is not that these procedures make reading the text *easier* or because text mining might "solve" a literary conundrum. In fact, such a reductive approach would be very uninteresting to the literary scholar. Many Stein critics argue that the "difficulty" the repetition engenders in *The Making of Americans* is valuable precisely because it is rooted in indeterminacy: that is, the text's use of repetition represents a postmodernist project that challenges readerly subjectivity and deconstructs the role language and the process of writing plays in determining meaning. The more "difficult" the text is to read, the more it meets its philosophical goals of making the reader question acts of representation and interpretation in general. Thus the text mining process is useful for literary analysis for three reasons:

1. Text mining may be used to determine relationships (clusters and patterns) in large bodies of data (in this case, the confusing network of Stein's repetitions).
2. Text mining can be used to identify relationships within the text that are not predetermined by meaning but are based on structural elements of the content (such as repetitive phrases or words).
3. Text mining also requires "subjective human evaluation" as an essential part of the analytical process.[3]

Analyzing *The Making of Americans* has already provided rich opportunities for thinking about both tool development and processes of literary analysis. For example, initial analyses on the text using the Data to Knowledge (D2K) application environment for data mining[4] (which was used in *the nora project*[5]) has yielded clusters based on the existence of frequent patterns. Instead of using single words for analyzing the text, the features used were phrases (i.e. ngrams). Executing a frequent pattern analysis algorithm[6] produced a list of patterns of the phrases co-occurring frequently in paragraphs. This frequent pattern analysis generated thousands of patterns because slight variations generated a new pattern. Because the number of frequent patterns was so large, another algorithm was applied that clustered these frequent patterns.[7] Although the analytics that were used are sophisticated, the results are not presented in a manner that makes them easy for scholars

to understand. For example, one cluster has 85 frequent patterns (see Figure 1). A scholar could piece clustered phrases back together to form a longer repetitive phrase ("part of Gossols where no other rich people were living"), but this is difficult, and, in addition, the results do not indicate where the phrase occurs in the document. The impact of Stein's repetition, after all, is always contingent on context, on what comes before and after a particular string of letters, on what those strings (or words) mean in the context of a sentence, a paragraph, and the larger narrative or project according both to Stein and to the scholar/user's theoretical perspective.

Thus, this first attempt at producing clusters was only marginally useful for scholarly study because the rules by which the items are clustered are not readily apparent and the results are detached from the context of the text. One method for helping scholars understand the rules is by including the scholar in identifying the features by which the frequent patterns are determined. For instance, she could develop a "productive" question about the text, a question that correlates to specific linguistic and narrative features that can be extracted and eventually mined. For example, a question might be: how do linguistic and syntactic variations reflect, advance, or deter from the progression of the text's more abstract features such as argument and narrative? Specifically, do changes in repetition correspond to the novel's evolving theories about identity and representation? And how? Features which might be considered include, but are not limited to, repetition across words and phrases (including "trivial" and "motivated" recurrences);[8] grammatical features (such as shifting perspectives, oscillating verb-tenses, or dropped pronoun referents); syntactic features (such as diagrammic versus run-on or fragmented sentences); and narrative features (such as shifts between description or "telling" and narrative or "showing"). More work needs to be done to create the features described above for use in the analytics. This paper will discuss our attempts to augment these analyses.

In addition, in order to link these patterns back to the context, it is essential to develop an environment that facilitates a self-reflective, critical evaluation of the scholar's theoretical perspective and ultimately, her ability to influence how data is culled, to assert what Ben Shneiderman has called "user control." Thus, the text's features must be visualized in a meaningful way so the scholar can "see" the text mining results and engage with the techniques to improve the results. For instance, in our first attempt, we created frequent patterns without using stemming. After reviewing the list of results, we realized that stemming was valuable to regulate pluralism, verb tense, etc. Stemming was deployed. Our experience with the *nora*interface has also demonstrated that providing easy access to context within the text is important to understanding the results of classification algorithms—a task which is relatively easy when the entire text is a short poem (see Figure 2). A scholar who is attempting to study repetition across the 900-page text, however, needs new ways to access the context of each instance of repetition. She needs tools that take her beyond searching tools which list the context of a single line (see Figure 3 for results using "grep" on a file). To this end, we have explored various visualization methods such as a weighted centroid (Figure 4), a scatterplot (Figure 5), a heat map (Figure 6), and a line graph (Figure 7).After our team considered the advantages and disadvantages of each of these tools, Anthony Don led the effort to design FeatureLens (Figure 8), a tool which allows the user to see features derived from text mining results within the context of the original text. This tool will serve as the main focus of this talk.

Ben Shneiderman calls the powerful combination of text mining and visualization, "discovery" based because it empowers users "to specify what they are seeking and what they find interesting."[9] Gertrude Stein has this to say about the same topic:

> It is very likely that nearly every one has been very nearly certain that something that is interesting is interesting them . . . The only thing that is different from one time to another is what is seen and what is seen depends upon how everybody is doing everything. This makes the thing we are looking at very different and this makes what those describe it make of it, it makes a composition, it confuses, it shows, it is, it looks, it likes it as it is, and this makes what is seen as it is seen.[10]

Text mining allows the literary scholar to "see" the text "differently" by facilitating analytical approaches that chart repetition across thousands of paragraphs and visualizations that empower her to tweak the results, focus her search, and ultimately (re)discover "something that is interesting." This paper will discuss how this process of analyzing repetition in *The Making of Americans* has impacted the ways in which *nora* has combined text mining and visualization applications.

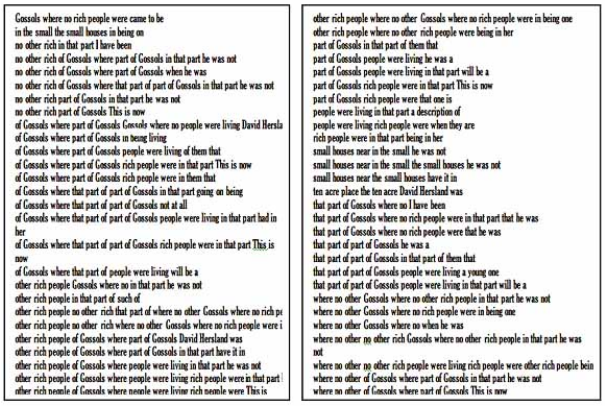


Figure 1: A list of frequent patterns for one cluster.

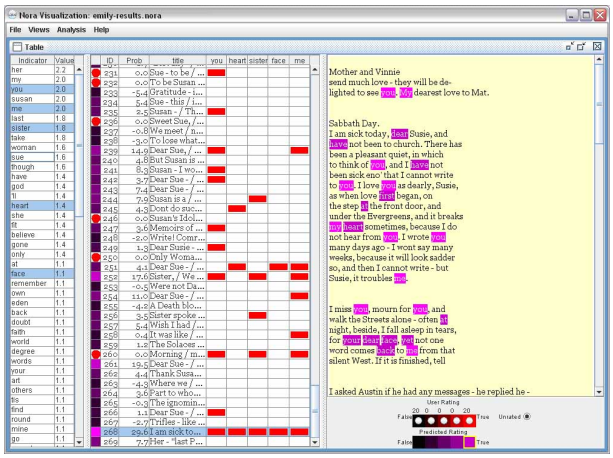


Figure 2: With the nora interface, users can review the results of the data mining, see which documents contain the feature words returned by the algorithm, and see the location of the words in a selected document. ()

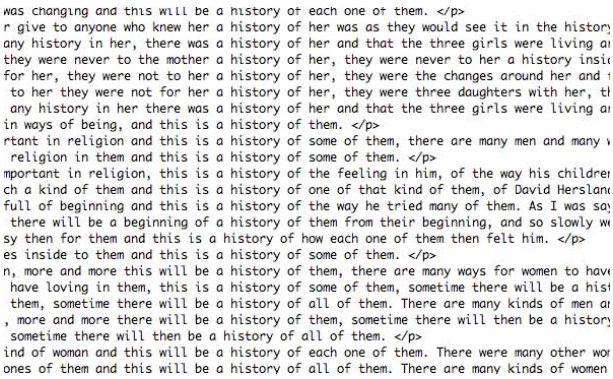


Figure 3: Identifying repetition as it emerges in a list of frequent pattern clusters in a simple "grep" text file.

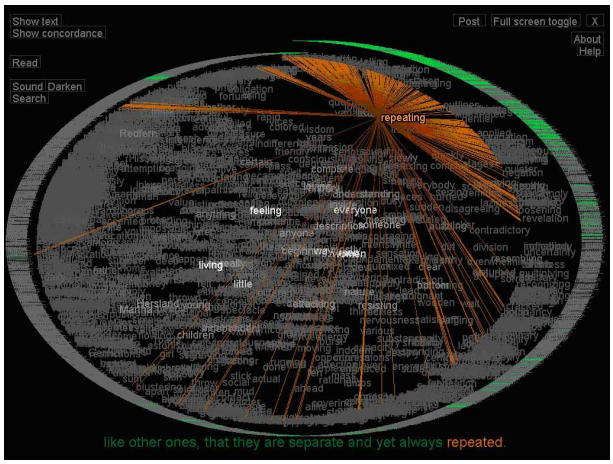


Figure 4: Using Text Arc, we are able to see where the word "Repeating" is repeated throughout the entire second section of The Making of Americans. The text of the section is represented by the lines in the outer ring. The word "repeating" is situated in the circle according to where it appears most often in the outer ring; the green lines represent lines in which the word appears; and the orange lines point to word's occurrence in those lines. ()

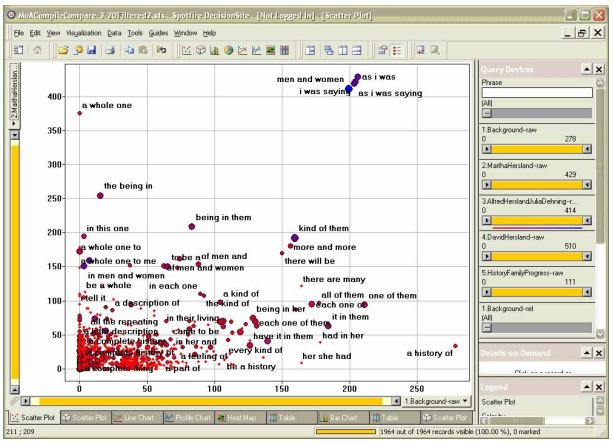


Figure 5: Common phrases displayed on a scatterplot, with frequencies in section 1 on the X axis and frequency in section 2 on the Y axis. We can see than "men and women" and "I was saying" is a lot more common than any other phrases, and used equally in both sections. ()

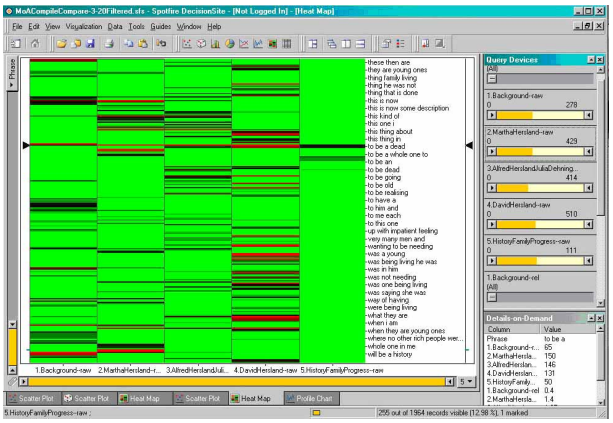


Figure 6: This heat map allows us to compare the frequency of phrases in five sections. Each line is a different phrase. The red lines show when a phrase occurs more than 100 times in a section. ()

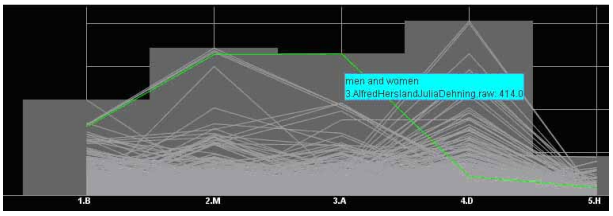


Figure 7: This line graph shows frequent phrases compared across five sections. Each line represents a different phrase. Here we can immediately see that "men and women" appears almost as frequently in two sections of the text. ()

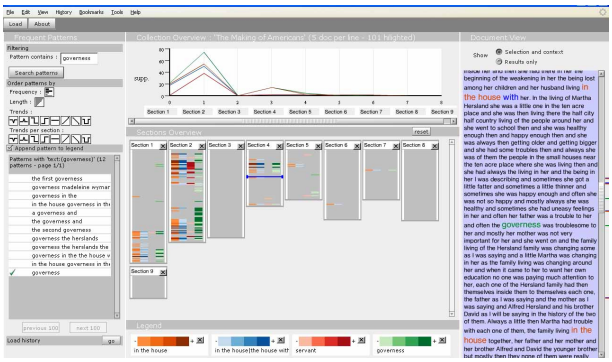


Figure 8: FeatureLens, created by Anthony Don.

1. Gertrude Stein, *The Making of the Americas: Being a History of a Family's Progress, (Normal, IL: Dalkey Archive Press, 1995). First published by Contact Editions, Paris, 1925.*
2. Malcolm Cowley, "Gertrude Stein, Writer or Word Scientist" *The Critical Response to Gertrude Stein*, (Westport, CT: Greenwood Press, 2000): 148.
3. This description of text mining is detailed further in Sholom M. Weiss et al., *Text Mining: Predictive Methods for Analyzing Unstructured Information*, (New York: Springer Science+Business Media, Inc., 2005)
4. Developed by the Automated Learning Group (ALG) at the National Center for Supercomputing Applications (NCSA), alg.ncsa.uiuc.edu
5. The nora project (www.noraproject.org) is a Mellon-funded collaborative (including computing, design, library science, andd English departments at the University of Alberta; University of Illinois, Urbana-Champaign; University of Maryland; University of Nebraska; and the University of Virginia) which is developing text mining and visualization software in order to "explore significant patterns across large collections of full-text humanities resources."
6. J. Pei, J. Han, R. Mao, "CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets (PDF)", *Proceedings of the 2000 ACM-SIGMOD International Workshop on Data Mining and Knowledge Discovery (DMKD'00), Dallas, TX, May 2000*.
7. Xifeng Yan, Hong Cheng, Jiawei Han, and Dong Xin, "Summarizing Itemset Patterns: A Profile-Based Approach," *Proceedings of the 2005 International Conference on Knowledge Discovery and Data Mining (KDD 05), Chicago, IL, August 2005*.
8. "Trivial" and "motivated" recurrences are used by Calvin Brown in his book *Repetition in Zola's Novels*, (Athens, GA: University of Georgia Press, 1952). Examples of such recurrences include "tags" (repeated descriptions such as "her face was worn, her cheeks were thin," "her worn, thin, lined determined face," "her lined, worn, thin, pale yellow face"), "key passages" (a relatively long repetition of a fundamental idea) and "hammer word" (a strong or exclusive obsession with any single idea).
9. Ben Shneiderman, "Inventing Discovery Tools: Combining Information Visualization with Data Mining." *Information Visualization* 1.1 (2002): 11.
10. Gertrude Stein, *Composition as Explanation*, (London: The Hogarth Press, 1926). Accessed 2006-11-01. <http://www.centerforbookculture.org/context/no8/stein.html>

Extending PhiloLogic

Charles M. Cooney
(cmcooney@diderot.uchicago.edu)
ARTFL Project
University of Chicago
Russell Horton *(russ@diderot.uchicago.edu)*
Digital Library Development Center
University of Chicago
Mark Olsen *(mark@barkov.uchicago.edu)*
ARTFL Project
University of Chicago
Glenn Roe *(glenn@diderot.uchicago.edu)*
ARTFL Project
University of Chicago
Robert L. Voyer *(rlvoyer@diderot.uchicago.edu)*
ARTFL Project
University of Chicago

This presentation will demonstrate how PhiloLogic 3[1] uses a mixed processing mode to leverage both the efficiency of relational databases and the richness of TEI-XML encoding. We will discuss the specific scripting and database techniques used to enable fast execution of object level searches on large collections of electronic texts. We will also show how this stand-off model can be extended to allow for more complex nested object and word searching.

By way of background and context, PhiloLogic is a full-text search, analysis, and retrieval tool developed by the ARTFL Project and the Digital Library Development Center at the University of Chicago. The main purpose of PhiloLogic is to run word searches on large corpora of literary texts and to display results quickly. The system supports full boolean and proximity searches and provides a number of reporting functions, including KWICs, collocation tables and a variety of word frequency breakdowns. To achieve the required speed, PhiloLogic pre-indexes each word, storing byte offsets in a flat file. PhiloLogic handles document structure, currently extending down to seven levels of depth, in a similar manner, abstracting a structure by assigning numbers to each nested object-level of each document. Thus, on the general continuum of database systems, from pure (or native) XML systems like eXist to traditional SQL systems like MySQL, PhiloLogic falls somewhere in the middle.[2] It is not just a word search engine. Rather, it uses an abstract representation of document structure, shredding the XML into sets of related database tables. This means that PhiloLogic can process TEI-XML or a variety of other document encoding schemes by extracting structural information from available text tagging. Storing both bibliographic data and object-level data in SQL tables, PhiloLogic can search document structure and refine word searching by using the shredded XML.

In the first part of our presentation, we will explain how PhiloLogic stores object-level data in MySQL for flexible and speedy results retrieval. PhiloLogic extracts metadata at various object depths from full-text documents and outputs that data into flat files. Called dividx.raw and subdividx.raw, these flat files contain object-level addresses and attributes of those objects. These flat files can be loaded into MySQL either manually, after PhiloLogic indexing finishes, or automatically, using the generated SQL scripts, load.database.sql and load.subdoctables.sql. When the user then executes a div-level search, the PhiloLogic search engine queries these tables, refining search results dynamically.

The user can thus limit searches to individual speakers, speaker genders, or other meta-characteristics, within a corpus of plays, a selection of acts, or even within scenes in individual works. Here is an example of dividx.raw from a database of plays:

5:1 Characters castlist 5

The first field contains the document id and the object-level separated by a colon. Here, counting up from zero, this is the first subobject in the sixth text. The second field contains the object header. The third denotes what kind of object it is, as specified in the type attribute of the div tag. The last field contains the document id again. The file subdividx contains data about lower level objects, such as individual speakers' speeches:

0:2:0:0:3 speaker CH00016 Lizette Grimaud F Businessman's wife French White Heterosexual 0

Once again, the first field contains the specific address of the object. The second field denotes that the object is a speaker tag. The rest of the fields contain metadata about the specific character. These stand-off tables allow the user to delimit word searches by speaker and speaker characteristics.

When applied to digitized dictionaries, this functionality allows the user to restrict searching to headwords, as opposed to article or entry bodies. A line from the divindex.raw from an Urdu dictionary:

0:1:538 अपुष्पति apushpit article अपुष्पति _apushpit 0 0

The first field, of course, is the object address for the headword. The following fields contain the headword in both romanized and UTF-8 formats.

In the second part of the presentation, we will show how PhiloLogic can be modified to talk to standoff concordance tables, not loaded into MySQL, that enable the user to search for alternate and variant forms of words. The indexing engine creates a flat, single-field concordance file, called words.R, containing the surface forms of every word in a database. After the load, the database administrator can run a simple script on this file to convert it into a multi-field table, called words.R.wom. By making a few standard edits to the word exploder, crapser, searches can be executed for simplified, lemmatized, or transliterated forms of each word. To cite an example from a collection of classical Greek texts, the words.R.wom file has fields that contain a non-accented Greek form of the word, the original form, and a romanized form:

συναγαπαν συναγαπ ν sunagapan

Any of these forms can be entered into the keyword search field to find the original, indexed word. On the search page, the user clicks a radio button to indicate which form she is entering. This selection triggers a subroutine in the word exploder that finds the entered form in words.R.wom, finally passing the indexed form to the search engine.

The extensions discussed above have been fully implemented to address the needs of particular projects. Looking forward, we have begun work on a new extension to the core PhiloLogic system to support several text mining tasks in conjunction with the standard full-text analysis capabilties. This effort, currently named "Philomine,"[3] will allow users to do perform text mining tasks on user selected subsets of a PhiloLogic database, facilitating comparative analyses of stylistic and content features at the document level. The first stage will use a set of freely avilable Perl modules, including SVMLight, naive-Bayesian, and decision tree modules. This will give the user a choice of algorithms when running data mining experiments. Result sets will be linked to the search functions of PhiloLogic, allowing for rapid inspection of text mining results. We hope that PhiloLogic extensions, coupled with these widely tested machine learning tools, will allow humanists to bring a new level of computational rigor to textual analysis and to broaden its scope, revealing differences and similarities across large textual corpora.

---

1. PhiloLogic is an open source software for Linux, OS-X, and Solaris systems used by a number of Digital Humanities projects. Please consult <http://philologic.uchicago.edu/> for additional information and recent releases.
2. For more on XML versus relational database systems, and options in between, see these articles: "Comparing XML and relational storage: A best practices guide," March 2005; Davies, Mark, "The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation", *International Journal of Corpus Linguistics*, Volume 10, Number 3, 2005, pp. 307-334; Lapis, George, "XML and Relational Storage – Are they mutually exclusive?", XTech 2005: XML, the Web and beyond, May 25-7, Amsterdam; Pardede, E., Rahayu, J.W., and Taniar, D., "Preserving Composition in XML Object Relational Storage", 19th International Conference on Advanced Information Networking and Applications (AINA 2005), 2, IEEE Computer Society, pp. 695-700, Taipei, Taiwan.
3. Suggested pronunciation: "Feelomeen," like the French feminine name.

The Anthropology of Knowledge: From Basic to Complex Virtual Communities in the Arts and Humanities

Stuart Dunn *(stuart.dunn@kcl.ac.uk)*
King's College London
Tobias Blanke *(tobias.blanke@kcl.ac.uk)*
King's College London

An online community may be defined very simply as any group of people who collaborate regularly and/or formally using internet technologies. Such communities are becoming increasingly important in arts and humanities research. They include blogs, wikis, mailing lists and other such fora. The Jiscmail and Listserv services (<http://www.jiscmail.ac.uk; http://www.lsoft.com>), and the availability of well-known open source wiki and blog software (e.g. <http://www.mediawiki.org/wiki/MediaWiki> and <http://twiki.org>) have made it so easy to establish online communities that most projects in the digital and humanities now have one of some kind or another. However such fora are founded on relatively basic technologies. The various arts and humanities e-science and cyberinfrastructure programmes underway in many countries offer huge opportunities to enable online communities in ways that go far beyond what is generally available now. These approaches will be based on incomparably more complex tools, methods and technologies such as Virtual Organizations (VOs), Virtual Research Environments (see <http://www.ahessc.ac.uk/briefing-paper>) collaborative virtual workspaces, and Semantic Web. If these tools, methods and technologies are to be exploited by practice-led arts researchers and humanists to the greatest effect, they will need to be grounded in a strategic, rigorous and systematic understanding of the behaviours of current online communities, as complex online systems will clearly evolve from collaboration tools currently in use. The proposed paper will offer an overview of current usage of various online fora, and propose a high-level mapping from that overview onto the capabilities of the e-science-based collaborative systems of the future, such as MyVocs (<http://www.myvocs.org/>).

Of immediate importance to the humanities are VREs, as they help geographically dispersed research groups to come together in virtual laboratories that allow modelling and experimentation as much as discussion of research results. The e-Science concept of a Virtual Organisation on the other hand is seen as a set of institutions and/or individuals defined by resource sharing policies. [1] A VO is a community with the will to share resources and information across the internet, while VREs enable VOs by adding tools and methods that help the community to work together and share resources in a secure manner. VREs attempt to bring together researchers across disciplines and administrative boundaries. Many argue that humanities computing in general constitutes such an attempt, where the 'glue' is to find 'methodological commons'[2] to present the disciplinary kinships among different disciplines in the humanities and in computing. A hi-tech VRE for arts and humanities computing could therefore make these computational methods the subject of online discussions enabled by advanced information discovery technologies. As a case study, this paper will report on our efforts to set up such a VRE.

The Arts and Humanities e-Science Support Centre (AHESSC) is in regular and direct contact with many online communities, and is in an excellent position to offer an overview. We will present findings from detailed research on the recent history of a range of existing discussion fora. These will include both moderated and unmoderated groups critically and qualitatively assessing the difference this makes, and extrapolate an agenda for the management systems needed for VREs. We will examine both direct and indirect links between messages, semantic commonalities, disciplinary vocabularies and threading patterns, and hyperlinking behaviour. We will also examine other well-established wiki and blog communities of relevance to the humanities in much the same way. The paper will draw out commonalities, identify points of conflict and the reasons underlying them, the implementation and practice of "netiquette' codes, and the importance of sustaining archives and maintaining access to them. We argue that regardless of the technology, these are essential questions if the adoption of advanced collaborative tools is to be a success. A key remit of AHeSSC itself is to help develop a community of e-science 'early adopters', and the research presented forms a central part of this mission. We believe that more complex collaborative technologies will not only enable the transmission of 'traditional' scholarly communication in the form of text and images as in the current basic systems; but also they will facilitate access to large scale, complex and fuzzy data, and allow scholars to work together in real time with that data. For example, instead of sending a complex dataset as an email attachment to a group whose members then manipulate it according to their expertise and locally available software, and then send it back to the originator, the new architectures will allow two or (many) more to work on it in real time, to discuss it, expose it to analytical tools, and to annotate it online, whilst preserving a complete record of the workflow. We will show how close critical analysis of the existing lo-tech systems should inform the design of such architectures.

This will be illustrated by considering a VRE for arts and humanities computing methodologies for which funding is currently being sought. This project will continue the community building efforts of the UK's AHRC ICT Methods Network (see <http://www.methodsnetwork.ac.uk>), with a taxonomy of computational methods developed by the Arts and Humanities Data Service (see <http://ahds.ac.uk/about/projects/documents/pmdb_taxonomy_v1_3_1.pdf>). This is seen as a first step to build a semantically enriched VRE, using the taxonomy as the foundation of an ontology. The taxonomy will be verified against papers from online databases or websites presenting tools and methods. Ontological approaches are used that allow not only the semantic integration of different text and multimedia resources, but also the tracking of exchanged arguments that help users better understand decisions about methodologies. Our paper will illustrate the evolutionary relationship between an advanced environment such as this, and the basic environments with which we are all familiar.

1. Foster and C. Kesselman, *The Grid 2: Blueprint for a new computing infrastructure*, (Morgan-Kaufmann, 2004).
2. See Willard McCarty *Humanities Computing*, (Basingstoke: Palgrave McMillan, 2005).

Open Source and Digital Humanities

Amy Earhart *(aearhart@tamu.edu)*
Department of English
Texas A&M University
Dominic Forest *(dominic.forest@umontreal.ca)*
École de bibliothéconomie et des sciences de l'information
Université de Montréal
James Smith *(jgsmith@tamu.edu)*
Computer Information Services
Texas A&M University

Funding agencies such as NEH and AHRC have clearly stated that the open source approach to digital humanities work is necessary for both short-term financial support of projects and long-term success of digital humanities work. However, both agencies have predominantly emphasized the creation of open source projects rather than asking scholars to consider the possibilities offered when projects tap into the use of externally produced open source software. Even less emphasis has been placed on how academic projects might structure projects to invite participation from the open source development community.

There are excellent examples of groups that have modeled the importance of open source methodologies in Digital Humanities including the TEI/XML movement to standardize an XML appropriate to humanities projects and NINES, currently developing open source tools for use by the broader community. Less common, however, is the exploration of how we might engage the developer community with our work. Examples abound in the larger technology framework. BBC's Backstage movement has created a model that asks a free developer community to produce vast numbers of prototypes based on existing RSS feeds and open source software. This movement provides a possible model and possible interfaces that could benefit digital humanities work. OSCON (Open Source Convention) offers another possible model. While academics struggle to find the appropriate tools for digital humanities and appropriate standardization practices within the context of academic restraints (budget and promotion issues in particular), we might consider how to engage the broader open source

developer community and how to rewrite and use open source software for our digital humanities projects.

This panel will explore various approaches to engaging with open source software and development. As is fitting for a panel that looks to engage a non-academic audience in the development of our projects, we have put together three papers from diverse contributors: a linguist interested in the use of open source software for text mining; a literary scholar interested in testing open source programs such as Ruby on Rails and Google maps, in a digital archive; and a computer programmer who has actively worked with the open source development community, and OSCON in particular, over the past 10 years. The panel participants will discuss the use of particular pieces of software, benefits and limitations of such approaches, ways to open digital humanities work to a broader development community, and how to engage the open source development community in our work.

Text Mining and Computer-Assisted Thematic Analysis: Using Open-Source Software to Discover Knowledge from Unstructured Documents

Dominic Forest

During the last decades, many humanities research initiatives have been concentrating their efforts on digitalizing large amounts of text (*British National Corpus*, *Brown corpus*, *Gallica project*, *Oxford Text Archive*, etc.). Simultaneously, technologies were developed for more accurate and faster digitalization processes, and for more flexible and robust digital document encoding (XML), storing and retrieval. In the last years, some results of these digitalization initiatives have been made available on various supports and in many digital formats.

Digitalization initiatives partly motivated the need for robust large-scale text analysis software.

The development of text analysis tools has been influenced (and has often integrated) concepts and techniques from many academic fields. Among the main disciplines involved in the conception and development of text analysis software, we find linguistics (natural language processing), information science, and computer science. However, in the last years important research developments in the general field of text analysis emerged from the areas of data analysis, artificial intelligence and machine learning. Initially, the concepts and techniques developed in these areas where mostly applied on numerical data. This led to a rapidly growing research area known as “data mining”. The primary objective of data mining tools is to rapidly extract valuable (and previously unknown) information from large sets of quantitative data. Many researchers, mostly familiar with the information retrieval community, saw the potential of this emerging technology in the domain of text analysis. The application of data mining concepts and techniques on large textual databases quickly became an active interdisciplinary research field known as “text mining”.

Although some more or less efficient commercial text mining software is becoming available, these tools suffer from two main problems in an academic research context. First, commercial text mining suites are very expensive and therefore have not been fully integrated in academic practices. Secondly, and more importantly, these suites have been conceived in order to mainly assist business-related text mining tasks. They are not adapted to the variety and specificity of humanities academic research. Nevertheless, very recent academic projects have started to explore the potential of text mining techniques applied to large humanities and social sciences corpora. These research initiatives are mainly based on open-source text mining modules.

The objectives of our talk are 1) to report an experiment in which we use open- source text mining tools to assist thematic analysis from unstructured texts and 2) identify the benefits and limits of using open-source text mining tools in the context of text analysis tasks in the humanities.

Our talk will be divided in three main parts. In the first part, we will present the fundamental concepts and techniques of text mining. More specifically, we will present how two fundamental knowledge discovery processes (text clustering and text categorization) can be assisted – and in some cases accomplished automatically – using state-of-the-art text mining techniques.

The second part will be dedicated to the presentation of a specific text mining methodology to assist thematic analysis of unstructured documents. This methodology is composed of four main steps: 1) text pre-processing, 2) text transformation using vector-space model, 3) knowledge discovery (using clustering or categorisation processes), and 4) thematic analysis. In this part, we will present each step in details and we will present open-source text mining tools that can accomplish each process.

In the third part, we will report the main results obtained using this four-step methodology on a Belgian newspaper corpus.

In conclusion, based on experiments using open-source text mining technology, we will identify and discuss some benefits and limits of open-source text mining tools in the context of humanities text analysis.

An Open-Source Approach to Digital Humanities: Testing the Limits of Open Source Ideology in the 19th-Century Concord Digital Archive

Amy Earhart

The 19th-Century Concord Digital Archive collaborators are exploring the utilization of a wide variety of open source tools in archive construction as well as following a distribution method that will encourage the participation of an open source development community. As the 19th-Century Concord Digital Archive nears beta test (Summer 2007), it seems an appropriate moment to consider the impact of the open source ideology of the project. A core value of many of the currently produced digital humanities projects is an open and free archive. This approach generally implies that users experience the free use of archive materials delivered through the web and, in rare instances, the development of open source tools that may be applicable to projects that use similar standards, such as TEI/XML. What has been given little attention, both in practice and in theory, has been a careful consideration of how non-academic open source programs might be leveraged for digital humanities projects, particularly digital archives, and what the positive and negative impact of such an approach might be. Using the Digital Concord project as a test case, the paper examines the possibilities offered by contemporary web-based technologies. This paper will discuss two such possible non-academic open source tools as well as planning practices. The presentation will then extrapolate from the experience of archive development to generalize some of the ways that non-academic open source approaches might positively and negatively impact digital humanities.

Of the various non-academic open source applications in use in the construction of the Digital Concord archive, this paper will focus on the two: Ruby on Rails and Google Map Hacks.

Ruby on Rails

Ruby on Rails has produced some of the hottest buzz of recent open source technologies. Rails is a web application framework that not only offers amazing capabilities, but has attracted some of the best and brightest of the open source development community. It is particularly useful to digital humanities work because it shows remarkable promise for rapid development of real world applications and works well with TEI/XML, relational databases, and RDF, a means of relating disparate information. And, given the interest in Rails, the development community has begun development of a broad set of beta programs that, with some modification, might prove helpful to our current digital humanities work. During the presentation, I will highlight some of the uses that Rails offers and discuss positive and negative impacts of the approach.

Google Map Hacks

In an attempt to provide multiple user interfaces and remedy the lack of responsiveness to a user experience with some digital humanities work, the Digital Concord project has sought to develop multiple interfaces. One such interface is visual and based on a manipulation of Google Maps. While GIS is a well-developed tool and has been applied to some digital humanities projects, there are other tools that might be of greater use to digital humanists due to cost, development, and specific disciplinary needs. Rather than rely on GIS tools, the Concord archive experiments with programs that are under development in an open source community in the expectation that such an approach will offer a way for academics to tap into a previously unexplored group of participants and developers. The advances made by individuals interested in such maps are ongoing; projects chronicled at the blog "Google Maps Mania" <http://googlemapsmania.blogspot.com/> show the possibilities for incorporating Google maps with a variety of data. Using the Concord interface as a starting point, the paper will discuss multiple possibilities for Google Map Hacks.

Designing Humanities Data for Open Source Developers

While the use of such open source technologies might have a very real impact on digital projects, the potential for engaging the open source community in digital humanities offers an even greater opportunity. Models from Moodle to the BBC backstage project <http://backstage.bbc.co.uk/> suggest that the possibilities for engaging a broad developer community in digital work is possible with the right set of data and appropriate interface. While the Digital Concord project has no illusions of generating the type of interest as seen in previously mentioned projects, there are ways that the project can open data and create a community of interest around digital humanities work that could have positive repercussions. The paper will close with a discussion of potential strategies that individual archives might take to interest the open source community.

Interfacing with the Open Source Community: An OSCOM approach

James Smith

Academic progress seems to depend on a person's ability to contribute to society as measured by attributable work such as articles and monographs in areas that are interesting to society. These attributions depend on copyrights and patents, the keys to establishing and protecting intellectual property ownership. The various print journals and publishing houses act as a peer-reviewed intermediary. The academy establishes social

interest in a project by the project's ability to attract financial support.

Digital Humanities (DH) relies heavily on what is now being called Web 2.0: technologies that allow nearly seamless cross-platform client/server applications (e.g., AJAX and Comet). Most of these technologies have roots in the Open Source Community (OSC). Much of the talent is also in the OSC. Until this talent is brought into DH, DH is in danger of being a field looking in from outside, watching technology advance as it tries to catch up. Yet, DH is rooted in the academy and must meet the expectations of the academy.

The Open Source Community has its roots in the academy, but has lived for a while in the wild among amateurs in the classic sense. Many people participate in open source because they enjoy doing so, though some participate because of their employment. The OSC is built on a stereotype of the academy: information is free and everyone is able to work on problems they find interesting. Just as professors enjoy working in a University or College with other smart people in their field, OSC members are attracted to projects with smart people. OSC members also value the openness with which they can develop and discuss projects.

The Open Source Community has not abandoned the notion of attribution and social relevancy. The openness of a project's development creates a trajectory along which the project travels. This trajectory is a measure of the talent behind the project and is apparent to most who are interested in the project. Misappropriating a particular release of a project captures only a single point on that trajectory. Because any particular release of a project does not bring with it any of the talent behind the project, `stealing' from the project is worth much less. Attribution in the OSC is much more than just the name beside the copyright or on the patent.

An Open Source project is socially relevant if it is widely used and has a strong community. There are no financial requirements for a project to be successful. Sorceforge.net hosts Open Source projects at no cost to the project not because any particular project is worth the cost, but because the OSC itself is socially relevant. People contribute their time to a project not necessarily because they get paid but because they enjoy the project and, if they make significant contributions, they can become a well known talent in the OSC.

The problem, then, is how to balance the requirements of the academy against the need to create an environment that is inviting to the open source community. By openly involving the open source community, DH can access a wide variety of talent which will be involved in various projects not because they are being paid to help, but because they love the project. At the same time, DH can maintain the attribution required for academic progress.

One academic field of study that is leading the way with technology is Physics with the development of the world wide web at CERN to aid in sharing documents to the electronic pre-print server (<http://xxx.lanl.gov/>). The Physics community can do this because it is small enough that its members know each other. Reputation is built much as it is in the Open Source Community: by personal experiences between members of the community. Formal peer review plays a secondary role. By reviewing the output of a physicist, another can see the pattern and tell if something new fits that pattern.

If any company represents the commercial potential of Digital Humanities, it is Google. They have been able to attract some of the top talent in the industry by providing a work environment that resembles the Open Source Community in many aspects. Some of the more interesting projects for DH have come from employees' `play time.' Google has brought a lot of smart people together under one roof, much as a University or College might do.

By looking to other academic fields and the Open Source Community, Digital Humanities can create a new environment encouraging rapid evolution of ideas without sacrificing the need for attribution, reputation, or social relevance.

Synergies: The Canadian Information Network for Research in the Social Sciences and the Humanities

Michael Eberle-Sinatra
(michael.eberle.sinatra@umontreal.ca)
University of Montreal

Funded by the Canada Foundation for Innovations, "*Synergies*: The Canadian Information Network for Research in the Social Sciences and Humanities" will be a national distributed platform with a wide range of tools to support the creation, distribution, access and archiving of digital objects such as journal articles. It will enable the distribution and use of Social Sciences and Humanities (SSH) research, as well as to create a resource and platform for pure and applied research. In short, *Synergies* will be a research tool and a dissemination tool that will greatly enhance the potential and impact of SSH scholarship.

Canadian SSH research published in Canadian journals and elsewhere, especially in English, is largely confined to print. The dynamics of print mean that this research is machine-opaque and hence invisible on the internet, where many students and scholars begin and sometimes end their background research. In bringing Canadian SSH research to the internet, *Synergies*will not only bring that research into the mainstream of worldwide research discourse but also it will legitimize online publication in SSH. The acceptance of this medium will open the manner in which knowledge can be represented. On one plane, researchers will be able to take advantage of an enriched media palette—color, image, sound, moving images, multimedia. On a second plane, researchers will be able to take advantage of interactivity. And on a third plane, those who query existing research will be able to broaden their vision by means of navigational interfaces, multilingual interrogation and automatic translation, metadata and intelligent search engines, and textual analysis. On still another plane scholars will be able to expand new areas of knowledge such as bibliometrics and technometrics, new media analysis, scholarly communicational analysis and publishing studies.

Canadian researchers in the SSH require two research communication services that can be provided within one structure. The first is an accessible online Canadian research record. The second is access to an online publication service that will place their work on record and will ensure widespread and flexible access. *Synergies* provides both these functions. Built on the foundation of *Érudit*, a Quebec-based research publication service provider in existence since 1998, *Open Journal Systems*, which is a tested online journal publishing software suite, and the technical expertise developed by its three other partners, *Synergies* will integrate work being done within its five-party consortium to create a decentralized national platform for SSH research communication. *Synergies* is designed to eventually encompass a range of formats—including published articles, pre-publication papers, data sets, presentations, electronic monographs—to provide a rich scholarly record, the backbone of which is existing and yet to be created peer-review journals. *Synergies* will bring Canadian SSH research into mainstream of worldwide research discourse using a cost-effective public/not-for-profit partnership to maximize knowledge dissemination.

The members of the *Synergies* consortium are the University of New Brunswick, Université de Montréal (lead institution), University of Toronto, University of Calgary, and Simon Fraser University. Each brings appropriate but different expertise to the project. At its first level, *Synergies* consists of this five-university consortium that will provide a fully accessible, searchable, decentralized and inclusive national SSH database of structured primary and secondary SSH texts. This distributed environment is technically complex to implement, and it also represents a major political and social collaboration which attests to the project's transformative dimension for Canadian SSH research and researchers. *Synergies* will be a primary aggregator of research that, in providing publishing services, will allow journals editors (and other producers) to structure subscriptions and maintain revenue control. At a second level, *Synergies* will reach out to 16 regional partner universities who will benefit from and contribute to extend *Synergies* functionality. At a third level, in a producer-to-consumer relationship with university libraries and organizations such as the Canadian Research Knowledge Network (CRKN), *Synergies* will make possible national accessibility. Using this relationship as a model, *Synergies* will be positioned to facilitate similar relationships for journals with buyer consortia around the world. *Synergies* is not only a pan-Canadian technical infrastructure but also a mobilizing and enabling resource for the entire scholarly community of Canadian SSH researchers. In embracing the whole of the social sciences and humanities, *Synergies* will foster cross-disciplinary, problem- and issue-oriented queries while also allowing queries that can be time-framed, discipline-based, media or methodologically specific, theoretically constrained or geo-referenced.

Synergies will also provide a needed infrastructure for the Social Sciences and Humanities Research Council (SSHRC) to follow through its in-principle commitment to open access and facilitate its implementation by extending the current venues

and means for online publishing in Canada. With *Synergies* in place the funding of journals based on dissemination effectiveness rather than sales levels will become both feasible for journals and possible as a evaluative criterion for SSHRC funding. The Canadian Federation for the Humanities and Social Sciences, with a membership of over 30,000, has also recently adopted a position in favor of open access and indicate the role that *Synergies* can play.

In summary, *Synergies* is the vehicle by which Canadian SSH research communication can be modernized. It embraces emerging research practice by utilizing existing texts, enriching, expanding, and greatly easing access to scholarly data and to audiences. It organizes a fragmented research record, ensuring and enhancing access to existing data sets. It facilitates access via aggregation of journals and an ability to facilitate agreements between Canadian SSH journals and other producers' and buyers' consortia such as CRKN. It lays a foundation for expanding the research record to encompass all scholarly inquiry in order to achieve maximum accessibility and circulation. *Synergies* represents a project in parallel with other national projects and disciplinary databases emerging in other countries, for example, *Project Muse*, *Euclid*, *JStor*, and *HighWire* in the United States, and in France, *Persée* and *Adonis*.

How Rhythmical is Hexameter: A Statistical Approach to Ancient Epic Poetry

Maciej Eder
(maciejeder@poczta.ijp-pan.krakow.pl)
Polish Academy of Sciences

In this paper, I argue that the later a given specimen of hexamater is, the less rhythmical it tends to be. A brief discussion of the background of ancient Greek and Latin metrics and its connections to orality is followed by an account of spectral density analysis as my chosen method. I then go on to comment on the experimental data obtained by representing several samples of ancient poetry as coded sequences of binary values. Hexameter's decreasing rhythmicity is then illustrated with reference to particular authors. In the last section, I suggest how spectral density analysis may help to account for other features of ancient meter.

The ancient epic poems, especially the most archaic Greek poetry attributed to Homer, are usually referred as an extraordinary fact in the history of European literature. For present-day readers educated in a culture of writing, it seems unbelievable that such a large body of poetry should have been composed in a culture based on oral transmission. In fact, despite of genuine singers' and audience's memory, epic poems did not emerge at once as fixed texts, but they were re-composed in each performance (Lord 2000, Foley 1993, Nagy 1996). The surviving ancient epic poetry displays some features that reflect archaic techniques of oral composition, formulaic structure being probably the most characteristic (Parry 1971: 37-117).

Since formulaic diction prefers some fixed rhythmical patterns (Parry 1971: 8-21), we can ask some questions about the role of both versification and rhythm in oral composition. Why was all of ancient epic poetry, both Greek and Latin, composed in one particular type of meter called hexameter? Does the choice of meter influence the rhythmicity of a text? Why does hexameter, in spite of its relatively restricted possibilities of shaping rhythm, differ so much from one writer to another some (cf. Duckworth 1969: 37-87)? And, last but not least, what possible reasons are there for those wide differences between particular authors?

It is commonly known that poetry is in general easier to memorize than prose, because rhythm itself tends to facilitate

memorization. In a culture without writing, memorization is crucial, and much depends on the quality of oral transmission. In epic poems from an oral culture rhythm is thus likely to be particularly important for both singers and hearers, even though they need not consciously perceive poetic texts as rhythmical to benefit from rhythm as an aid to memory.

It may then be expected on theoretical grounds that non-oral poems, such as the Latin epic poetry or the Greek hexameter of the Alexandrian age, will be largely non-rhythmical, or at least display weaker rhythm effects than the archaic poems of Homer and Hesiod. Although formulaic diction and other techniques of oral composition are noticeable mostly in Homer's epics (Parry 1971, Lord 2000, Foley 1993, etc.), the later hexameters, both Greek and Latin, also display some features of oral diction (Parry 1971: 24-36). The metrical structure of hexameter might be quite similar: strongly rhythmical in the oldest (or rather, the most archaic) epic poems, and less conspicuous in poems composed in written form a few centuries after Homer. The aim of the present study is to test the hypothesis that the later a given specimen of hexameter is, the less rhythmical it tends to be.

Because of its nature versification easily lends itself to statistical analysis. A great deal of work has already been done in this field, including studies of Greek and Latin hexameter (Jones & Gray 1972, Duckworth 1969, Foley 1993, etc.). However, the main disadvantage of the methods applied in existing research is that they describe a given meter as if it were a set of independent elements, which is actually not true. In each versification system, the specific sequence of elements plays a far more important role in establishing a particular type of rhythm than the relations between those elements regardless their linear order (language "in the mass" vs. language "in the line"; cf. Pawlowski 1999).

Fortunately, there are a few methods of statistical analysis (both numeric and probabilistic) that study verse by means of an ordered sequence of elements. These methods include, for example, time series modeling, Fourier analysis, the theory of Markov chains and Shannon's theory of information. In the present study, spectral density analysis was used (Gottman 1999, Priestley 1981, etc.). Spectral analysis seems to be a very suitable tool because it provides a cross-section of a given time series: it allows us to detect waves, regularities and cycles which are not otherwise manifest and open to inspection. In the case of a coded poetry sample, the spectrogram shows not only simple repetitions of metrical patterns, but also some subtle rhythmical relations, if any, between distant lines or stanzas.

To verify the hypothesis of hexameter's decreasing rhythmicity, 7 samples of Greek and 3 samples of Latin epic poetry were chosen. The specific selection of sample material was as follows: 3 samples from Homeric hexameter (books 18 and 22 from the Iliad, book 3 from the Odyssey), 1 sample from Hesiod (Theogony), Apollonius (Argonautica, book 1), Aratos (Phainomena), Nonnos (Dionysiaca, book 1), Vergil (Aeneid, book 3), Horace (Ars poetica), and Ovid (Metamorphoses, book 1). In each sample, the first 500 lines were coded in such a way that each long syllable was assigned value 1, and each short syllable value 0. Though it is disputed whether ancient verse was purely quantitative or whether it also had some prosodic features (Pawlowski & Eder 2001), the quantity-based nature of Greek and Roman meter was never questioned. It is possible that rhythm was generated not only by quantity, but it is certain that quantity itself played an essential role in ancient meter. Thus, in the coding procedure, all prosodic features were left out except the quantity of syllables (cf. Jones & Gray 1972, Duckworth 1969, Foley 1993, etc.). A binary-coded series was then obtained for each sample, e.g., book 22 of the Iliad begins as a series of values:
1110010010010011100100100100111001110010010011...

The coded samples were analyzed by means of the spectral density function. As might be expected, on each spectogram there appeared a few peaks indicating the existence of several rhythmical waves in the data. However, while the peaks suggesting the existence of 2- and 3-syllable patterns it the text were very similar for all the spectograms and quite obvious, the other peaks showed some large differences between the samples. Perhaps the most surprising was the peak echoing the wave with a 16-syllable period, which could be found in the samples of early Greek poems by Homer, Hesiod, Apollonius, and Aratos. The same peak was far less noticeable in the late Greek hexameter of Nonnos, and totally absent in the samples of Latin writers. Other differences between the spectograms have corroborated the observation: the rhythmical effects of the late poems were, in general, weaker as compared with the rich rhythmical structure of the earliest, orally composed epic poems.

Although the main hypothesis has been verified, the results also showed some peculiarities. For example, the archaic poems by Homer and Hesiod did not differ significantly from the poems of the Alexandrian age (Apollonius, Aratos), which was rather unexpected. Again, the rhythm of the Latin hexameter turned out to have a different underlying structure than that of all the Greek samples. There are some possible explanations of those facts, such as that the weaker rhythm of the Latin samples may relate to inherent differences between Latin and Greek. More research, both in statistics and in philology, is needed, however, to make such explanations more nuanced and more persuasive.

Bibliography

Duckworth, George E. *Vergil and Classical Hexameter Poetry: A Study in Metrical Variety.* Ann Arbor: University of Michigan Press, 1969.

Foley, John Miles. *Traditional Oral Epic: "The Odyssey", "Beowulf" and the Serbo-Croatian Return Song.* Berkeley: University of California Press, 1993.

Gottman, John Mordechai, and Anup Kumar Roy. *Sequential Analysis: A Guide for Behavioral Researchers.* Cambridge: Cambridge University Press, 1990.

Jones, Frank Pierce, and Florence E. Gray. "Hexameter Patterns, Statistical Inference, and the Homeric Question: An Analysis of the La Roche Data." *Transactions and Proceedings of the American Philological Association* 103 (1972): 187-209.

Lord, Albert B. Ed. Stephen Mitchell and Gregory Nagy. *The Singer of Tales.* Cambridge, MA: Harvard University Press, 2000.

Nagy, Gregory. *Poetry as Performance: Homer and Beyond.* Cambridge: Cambridge University Press, 1996.

Parry, Milman. Ed. Adam Parry. *The Making of Homeric Verse: The Collected Papers of Milman Parry.* Oxford: Clarendon Press, 1971.

Pawlowski, Adam. "Language in the Line vs. Language in the Mass: On the Efficiency of Sequential Modeling in the Analysis of Rhythm." *Journal of Quantitative Linguistics* 6.1 (1999): 70-77.

Pawlowski, Adam, and Maciej Eder. "Quantity or Stress? Sequential Analysis of Latin Prosody." *Journal of Quantitative Linguistics* 8.1 (2001): 81-97.

Priestly, M. B. *Spectral Analysis and Time Series.* London: Academic Press, 1981.

From TEI to a CIDOC-CRM Conforming Model: Towards a Better Integration Between Text Collections and Other Sources of Cultural Historical Documentation

Øyvind Eide *(oyvind.eide@muspro.uio.no)*
Unit for Digital Documentation
University of Oslo
Christian-Emil Ore *(c.e.s.ore@edd.uio.no)*
Unit for Digital Documentation
University of Oslo

Introduction

In the last couple of years, there has been a growing interest towards including into TEI documents information about the world rather than information concerning the text of the document to be encoded only. We see examples of this through recent additions to the TEI standard, e.g. the person element (TEI P5, sec. 20.4.2), as well as through the work in the Ontologies SIG since it was established in 2004 (TEI Ontology SIG WIKI). In the SIG, the topic of discussion is how to organise this kind of information about the world according to specific ontologies.

One particularly promising ontology in this context is the CRM (CIDOC 2003). It has been used together with Topic Maps to organize information from TEI documents (Tuohy 2006) and as an attempt to find a solution to the so-called exhibition problem (Eide 2006). Further, attempts have been made to formalise a way to connect TEI and CRM documents (Ore 2006).

In this paper, we propose a method for automatic generation of CRM conforming models based on TEI documents. We will discuss limitations to this approach, as well as ways these may be overcome.

Our proposed method

The method we propose will include two important steps that should be possible to implement to any given TEI document: Mapping and model building.

Mapping

A mapping from the TEI document into a model conforming with CRM should be created. It will be based on a general mapping of TEI elements to CRM we are currently developing. But in TEI, many elements are defined quite loose, and depending on the way they are used, they may be modelled differently in CRM. According to the TEI guidelines, tag usage may be described in the TEI header. Such descriptions may help to decide which type of modelling is the most appropriate.

Ideally, such a specific mapping should be created based on an automatic reading of the TEI header. But an element description in a *tagUsage* element in the TEI header is in prose and will generally not be stringent enough to be understood by an automatic reading (TEI P5, sec. 5.3.4). Human interaction will be needed. It may be the case that use of the *equiv* element will make automatic creation of mappings possible, as a reference to a certain CRM class may be included as an external link (ibid, sec. 6.3.4).

Model building

A CRM conforming model based on the TEI document and populated with all instances of mapped elements should then be created. This model may be used as a query or a data mining system where the user looks for interesting structures in the CRM conforming model alone, as well as in combination with textual information collected from the TEI source document. But this model may also be used in connection with other CRM conforming models, such as museum databases. The connections will be based on regional or global object identification, such as authority lists of names and classification schema. The resulting "super model" may then be used as a data mining tool based on semantic integration between heterogeneous resources.

An example of a TEI input document

We are currently developing the building blocks for a system based on the method described above. The example text used in our work is taken from a manuscript describing examinations about geographical matters performed as court interviews in 1740s, printed in the 1960s (Schnitler 1962). The printed edition was digitised and marked up in a typical TEI way, with names of places and people as well as dates tagged. A short paragraph of this tagged version, based on page 73 in the printed book, is translated to English and included below.

 Answ: Named <name type="person" xml:id="s1_24454"> Ole Nilsen</name>, is born in <name type="place" xml:id="s1_24457"> Tydals </name> mountains, Which is in
<name type="place" xml:id="s1_24460"> Norway </name>, of Sami parents, is 50 years old, married, and having one child; has mostly dwelled in <name type="place" xml:id="s1_24466" > Tydals </name> mountains, and now dwelling in the Norwegian <name type="place" xml:id="s1_24469"> Mærragers </name> mountains.

What is tagged — and what is not

Names and dates are tagged in our document. This means that many references to persons, places and times are included. But there are a lot of other references to similar real world entities that are not tagged, e.g. when words other than proper names are used to refer to them. In the example above, "one child" is not tagged, whereas other references to the same historical person in terms of his or her name are tagged as person names. Furthermore, events are not tagged in this text. Thus, whereas the name of a boy who is born and the place of his birth is tagged, the event connecting these together, i.e. the birth, is not marked up.

This is common to most TEI documents, and is based on the text centric tradition of the TEI community. There were good reasons why this tradition was established, but it may be the case that for some types of documents, textual references to e.g. events should be marked up.

A simple CRM model of parts of the paragraph above is included as Figure 1. The solid lines represent what can be directly read from XML elements in the TEI document, whereas the dotted lines shows the parts not based on the TEI markup. This shows that the information needed to model the connection between the person and the mountain - that he was born there - is not tagged in the TEI document.

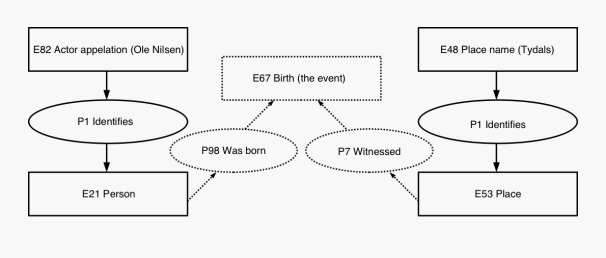


Figure 1

How to find the events

As the models created through the use of our proposed method is based on the XML tags alone, important information in the text is not considered when the model is created. Several possible ways to improve this exist. One is automatic event detection, as used e.g. in the Perseus project (Smith 2002). The use of this method too causes some problems. One problem is that fact that even if the system works quite good for English material, considerable work remains to be done to create a similar tool for 18th century Danish, even though the method is implemented also for smaller languages such as Finnish (Makkonen 2003). Named entity recognition systems are developed for modern Danish (Johannessen 2004), but this is of no use in the common situation where names are already tagged in the documents. Another problem is the fact that these kinds of methods will always be imperfect, resulting in either missed events, false positives, or both of these.

Another way to solve the problem would be to reread the text, identifying events and tagging them. This would be a reliable method, but time-consuming. Even a combination of these strategies, a semi automatic method, would mean quite a lot of work.

A possible way to solve the problem of event identification is to use the model. Any person or place in the CRM model has a link to the name of the person or place, an from the name in the CRM model to the TEI name element. This means that the textual distance to other person name elements, place name elements and date elements can be calculated. Further, it is possible to locate all words within a certain distance, and all words between two names.

This is similar to the first approach above. But in addition, we propose to connect the CRM model to external databases. An example of this would be if we have a CRM version of a database based on church book records. In this church book based CRM model a person may be found with a name similar to Ole Nilsen, a birth date in a possible range for being a witness in 1742, and a birth place with a name similar to a place name mentioned in the text in connection to the person's name. This external source may then help us to include the E67 Birth event in our model. This may turn out to be impossible without manual work, but we hope at least to make the manual work more effective.

Conclusion

A general observation from our work is that the more relevant information types is marked up in a TEI document, the easier it is to use automatic methods to generate CRM conforming models. But even a limited tagging with only names and dates marked up do help. We will continue our work on the implementation of a system based on the method described in this abstract. We believe this will improve the usability of TEI documents as information sources as well as simplifying the process of adding more information, such as event elements, into such documents.

Bibliography

TEI Ontology SIG WIKI. Accessed 2006-11-12. <http://www.tei-c.org.uk/wiki/index.php/SIG:Ontologies>

Burnard, Lou, and Syd Bauman, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Ver. 0.5. Accessed 2006-11-13. <http://www.tei-c.org/release/doc/tei-p5-doc/html/>

CIDOC. "Definition of the CIDOC Conceptual Reference Model." ISO/DIS 21127. 2003. Accessed 2006-11-13. <http://cidoc.ics.forth.gr/definition_cidoc.html>

Eide, Øyvind. "The Exhibition Problem. A Real Life Example with a Suggested Solution." *Digital Humanities 2006 Conference Abstracts.* Paris: CATI, Université Paris-Sorbonne, 2006. 58-61.

Johannessen, Janne Bondi, Eckhard Bick, Kristin Hagen, Dorte Haltrup, Åsne Haaland, Andra Björk Jónsdottir, Dimitrios Kokkinakis, Paul Meurer, and Anders Nøklestad. "The Nomen Noscio Project - Scandinavian Named Entity Recognition." *ALLC/ACH 2004 Conference Abstracts.* Göteborg: Göteborg University, 2004.

Makkonen, Juha, and Helena Ahonen-Myka. "Extraction of Temporal Expressions from Finnish Newsfeed." *Proceedings of 14th Nordic Conference of Computational Linguistics (NoDaLiDa 2003).* Reykjavik, 2003.

Ore, Christian-Emil, and Øyvind Eide. "TEI, CIDOC-CRM and a Possible Interface between the Two." *Digital Humanities 2006 Conference Abstracts.* Paris: CATI, Université Paris-Sorbonne, 2006. 62-65.

Schnitler, Peter. *Major Peter Schnitlers grenseeksaminasjonsprotokoller 1742-1745. Bind 1 [Major Peter Schnitler's border examination protocols 1742-45. Volume 1].* Oslo, 1962.

Smith, David A. "Detecting Events with Date and Place Information in Unstructured Text." *International Conference on Digital Libraries. Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries.* Portland, 2002. 191-196.

Tuohy, Conal. "Topic Maps and TEI – using Topic Maps as a tool for presenting TEI documents." *TEI Day in Kyoto 2006.* 2006. 85-98.

Bits and Pieces of Text: Appraisal of a Natural Electronic Archive

Maria Esteva *(mesteva@mail.utexas.edu)*
School of Information
University of Texas at Austin

This paper presents the methods and tools designed to appraise a digital archive. The attributes characterizing the archive at hand led to the development of the concept of *natural electronic archives* that would allow transforming the archive as a whole into a unit of analysis. The methodology for appraising the archive combines traditional archival concepts while adding tools –such as text mining and social network analysis– taken from other fields.

The digital archive belongs to a philanthropic agency whose activities in support of the arts, sciences and social welfare in Argentina span from the mid 1980s until its closure in December of 2005. In the early 1990s the institution implemented a networked server to store, retrieve, and share electronic texts, email, applications, and databases. In it, each employee had a virtual folder with his or her initials to store the files that they created. To comply with legal requirements the archive must be kept closed but accessible for the next 10 years. In parallel, appraisal needed to be undertaken to decide on the archive's long term destiny. An initial survey of the archive led to the development of the natural archive concept and suggested an appraisal method to establish archival evidence.

The natural electronic archive concept grew while I was surveying the server and from the interviews conducted with staff members who had worked in the organization since it opened. Based on my analysis, I concluded that the manner in which records were generated and kept could not easily be ascribed to digital archiving models currently discussed in the literature. These models focus more on the creation of sound electronic records, the design of electronic record-keeping systems, and on institutional repository archiving models than on the way in which digital archives exist "in the wild" (Bearman & Trant, 1997; Cox, 1997; Duranti et al., 2002; InterPARES Authenticity Task Force 2002; Jones et al., 2006). The natural archive concept builds partly on that of "natural collections" proposed by Phillip Cronenwett to describe collections of literary manuscripts as they leave the hand of the creator (Cronnenwett, 1984, p.106). I suggest that this concept

is relevant both to the case study at hand, as well as to archives of public or private persons and institutions showing similar characteristics.

Creation of natural electronic archives involves a set of *ad-hoc* practices developed as people adjust to and learn how to use information technologies. A natural archive is not designed or managed by records managers or archivists. Instead, it is what those working in institutions, in different capacities, using different technologies, and making decisions, make of it. In a natural archive, records are created, named, destroyed, or retained according to individual work-practices. Each record creator decides on structure and naming conventions for files and folders, spontaneously or consistently, according to individual mnemonic rules or the spur of the moment. Within the virtual folders, images, spreadsheets, texts, web sites, databases, back-ups, and applications live together under the same roof, placed or misplaced, in organized or disorganized fashion, in general without descriptive clues.

In a context without explicit recordkeeping rules, bits and pieces of text are ubiquitous inhabitants. Either shared by different members of a network or used repeatedly by their creators, they constitute the core of many records. This repetition of fragments afforded by the cut and paste function of the text editor, speaks as much of provenance, group collaboration and fair use, as of hierarchies and corporate culture. As a consequence of all of these phenomena, records within a natural archive are difficult to identify and lack formal documentation. This creates doubts about their capacity to provide evidence.

An appraisal method that uses Text Mining and Social Network Analysis was designed to determine what type of evidence of the organization that created it is provided by a natural electronic archive. The method is rooted in concerns expressed by Peter Boticelli in his study of networked organizations (Boticelli, 2000). It considers the need to document dynamics and changes in organizations and it explores the meaning of evidence and archival bond –understood as the "network of relationships between records" – in an ambiguous environment (Duranti & Guercio, 1997). Also, it highlights the importance of preserving evidence of the archive's formation process that will allow the study of its technological history and social uses (Lubar, 1996; Parezo, 1996). Its main departure from other appraisal methods is that it uses digital tools to analyze a large corpus of records inductively.

Determining archival evidence implies being able to map the organization through its records. Text Mining and Social Network Analysis use computing algorithms to discover knowledge about the relations among electronic records. By measuring the similarity between texts produced and co-produced by staff members within frameworks of time and provenance, the strength of relationships between records and between the staff members and/or functions that created them can be established. In turn, by averaging the similarities between the records of all pairs of staff members or functions across time, organizational structure and functions as well as correspondent changes in dynamics emerge. To confirm the validity of the findings, results are contrasted against the narratives of staff members about who they collaborated with, when, and in what. In this way, the evidence provided by the electronic records in this natural archive can be attested

A proof of concept was conducted to determine the feasibility of the appraisal method. For this, copies of electronic text records from the archive were used, while the original archive, kept with its directory structure intact, remains as guarantee of provenance and original order. Pre-processing documents involved using file management software to sort files within directories and sub-directories to construct sets belonging to a group of staff members that worked in the organization during a one year period. After conversion to .txt format, the sets were submitted to Rainbow, an open source text classification and retrieval tool, to obtain a vector space model (McCallum, 1996). In this phase, several trials were conducted to find the best way to narrow the vocabulary without losing language subtleties. From this model, pair-wise distances between documents where calculated using the cosine similarity formula in MatLab on a UNIX server. The resultant matrix was submitted to the social network analysis software UCINET to obtain a network drawing of the distances between texts (Borgatti et al., 2002). In turn, the average of distances corresponding to each staff member were calculated to obtain a matrix of relationships between staff members during one year.

The experiment suggested that relationships between staff members do emerge from the similarities and differences between the texts that they create. Testing showed that staff members who were leaving the organization and wrote farewell or personal records were less related to those cooperating in the preparation of monthly or annual reports. Also, project proposals written by grant applicants and stored in the shared server were barely related to reports or appropriation requests written by members of the organization. These preliminary results indicated that shifts in functions and consequent relationships between staff members can emerge from electronic texts.

The proof of concept also examined the use of cluster analysis to explore the concept of archival bond in natural electronic archives. Analyzing the content of strongly and poorly related records can explain what characterizes relationships between records –provenance, date, type of record, contents, topic – and whether these features can be mapped onto theoretical conceptualizations of archival bond. It will also explain the role of drafts, versions, and non-records by finding the proportion in which they exist in the natural archive and how close or not they are to complete records.

Before issuing the final appraisal protocol, changes had to be implemented and concerns addressed. Since the archive contains formats as old as Microsoft Word 5.0 for DOS, a converter with broad file format support was found to transform old files to ANSI text so they can be processed by Rainbow. Because there are various pieces of software involved, processes need to be automated and simplified as much as possible and issues related to the size of the text sets and matrices vis a vis the power of the processing tools have to be considered. Through a research grant from the University of Texas at Austin a programmer was hired to modify existing applications and develop new ones. Rainbow's tokenizer was modified to recognize Spanish characters and to include the Oleander Spanish stemmer (Oleander Solutions, 2006). The cosine similarity algorithm was coded in C++ so that bigger matrices can be processed efficiently in a UNIX server. To improve the ability to distinguish the characteristics of individual texts, Tf-idf capabilities were added to the script. The program outputs both a matrix of cosine similarities between every other document and a matrix of the averages of cosine similarity distances between every other author in the sample. Current testing involves processing sets with all the texts produced in one year by every author to determine changes in collaboration dynamics. After processing the matrices with UCINET, preliminary results for author's yearly averages show relationships that concur with their functions in the institution (See Fig. 1 and 2).

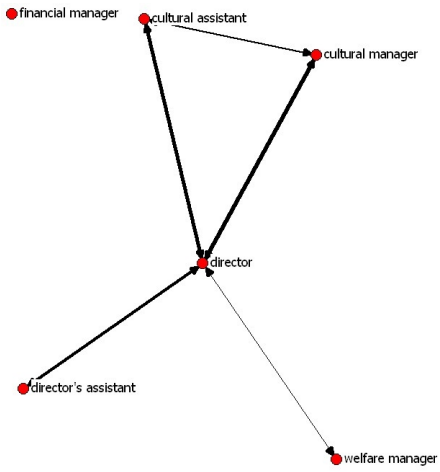


Figure 1. Network drawing of averages of cosine similarities between a set of 544 records of different staff members during the year 1996. The director is at the center of the network which corresponds with his functions in the organization and the data gathered in the interviews. Most of the records produced by the financial manager are non-textual and remained within the database systems which explains the distance between him and the director.

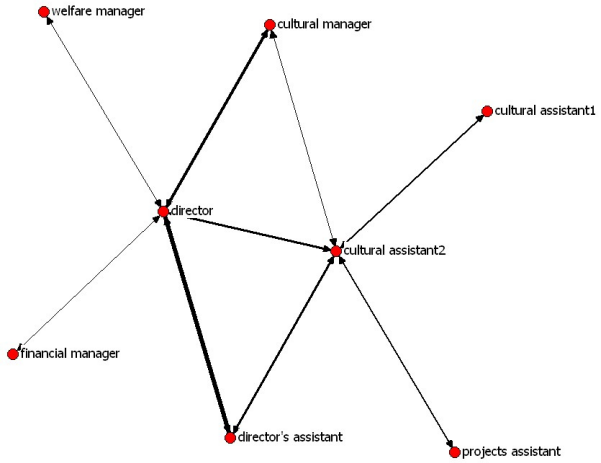


Figure 2. Network drawing of averages of cosine similarities between 719 records of different staff members during the year 1997. As new staff members added their records in the networked directory and started their own relationships, the majority of the director's previous relationships and his status in the network remained stable.

The use of Text Mining and Social Network Analysis promises to allow archivists to explore and define the meaning of evidence in natural electronic archives. Instead of intuition and art, as appraisal has been characterized, the opportunity exists to use inductive quantitative methods to make of appraisal a research endeavor (Eastwood, 1992). Moreover, the use of these methods opens the doors to the enormous potential of digital tools in the analysis and processing of digital archives.

Bibliography

Authenticity Task Force. *Requirements for Assessing and Maintaining the Authenticity of Electronic Records.* InterPARES, 2002. Accessed 2005-10-06. <http://www.InterPARES.org/display_file.cfm?doc=ip1_authenticity_requirements.pdf>

Bearman, David, and Jennifer Trant. "Electronic Records Research Working Meeting May 28-30, 1997: A Report from the Archives Community." *D-Lib* (July/August 1997). Accessed 2005-10-03. <http://www.dlib.org/dlib/july97/07bearman.html>

Borgatti, Stephan P., , and Freeman, L.C. UCINET 6 Social Networks Analysis Software. 2002. Accessed 2005-04-01. <http://www.analytictech.com/ucinet/ucinet.htm>

Boticelli, Peter. "Records Appraisal in Network Organizations." *Archivaria* 49 (2000): 161-191.

Cox, Richard J. " Electronic Systems and Records Management in the Information Age: An Introduction." *ASIS* 23.5 (1997).

Accessed 2005-10-03. <http://www.asis.org/Bulletin/Jun-97/cox.html>

Cronenwett, Philip N. "Appraisal of Literary Manuscripts." *Archival Choices: Managing the Historical Record in an Age of Abundance.* Ed. Nancy E. Peace. Lexington, MA: Lexington Books, 1984. 105-116.

Duranti, Luciana, Terry Eastwood, and Heather McNeil. *Preservation of the Integrity of Electronic Records.* Boston: Kluwer Academic, 2002.

Duranti, Luciana, and Maria Guercio. "Research Issues in Archival Bond." Electronic Records Meeting, Session I. 1997. Accessed 2006-03-29. <http://www.archimuse.com/erecs97/s1-ld-mg.HTM>

Eastwood, Terry. "Towards a Social Theory of Appraisal." *The Archival Imagination: Essays in Honor of Hugh A. Taylor* . Ed. Barbara L. Craig. Ottawa: Association of Canadian Archivists, 1992. 71-89.

Jones, Richard, Theo Andrew, and John MacColl. *The Institutional Repository*. Oxford, UK: Chandos Publishing Limited, 2006.

Lubar, Steven. "Learning from Technological Things." *Learning from Things.* Ed. W. David Kingery. Washington, D.C.: Smithsonian Institution Press, 1996. 31-34.

McCallum, Andrew K. Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering. 1996. Accessed 2005-02-02. <http://www.cs.cmu.edu/~mccallum/bow>

Oleander Solutions. Oleander Stemming Library. 2006. Accessed 2006-11-24. <http://www.oleandersolutions.com/stemming.html>

Parezo, Nancy J. "The Formation of Anthropological Archival Records." *Learning from Things.* Ed. W. David Kingery. Washington, D.C.: Smithsonian Institution Press, 1996. 145-174.

Rushdie's Computers: Born-Digital Archives and Humanities Scholarship

Erika Leigh Farr *(elfarr@emory.edu)*
Emory University

As questions and concerns about digital preservation and sustainability become increasingly audible in the spheres of digital humanities and humanities computing, the necessity to build strong ties between digital humanities and digital libraries only intensifies. Howard Besser, in his essay, The Past, Present, and Future of Digital Libraries," underscores the connections between these two fields. He explains digital libraries not only "provide access to original source material, contextualization, and commentaries, but they also provide a set of additional resources and service".[1] Besser then delineates some of these resources and services, including digital collections of traditional print materials, lexical analysis, and increased accessibility. In addition to these important contributions to humanities research, I would like to highlight the emerging role libraries play in processing, preserving, maintaining, and providing access to important archives that house born-digital content. This role will not only contribute to humanities research in decades to come, but will also impact how research is performed by directing what content is made available and how researches may access it.

In this paper, I will examine the Emory University Libraries' acquisition and processing of a singular personal archive as a case study to explore the methods and practices of handling born-digital archival materials and the implications such methodologies and their outcomes may have on humanities research.

Emory University's acquisition of Salman Rushdie's personal archive represents an important addition to the Manuscript and Rare Books Library, and contributes significantly to the University's digital library resources and research. Rushdie's rich personal archive includes traditional manuscript materials such as journals, personal correspondence, and notebooks, as well as less traditional archival materials, namely a series of personal computers that cover a significant span in his personal and literary life. This digital archive includes five computers, one early Macintosh desktop and four Macintosh laptops, including both obsolete and current models. While MARBL has previously acquired collections containing some digital materials, Rushdie's computers represent the first significant,

sizeable digital component to the University's extensive holdings of rare and unique materials.

Such an acquisition requires archivists to engage with technologists to ensure that the library can most effectively serve current and future researchers and scholars. The curation of such an archive raises important questions about how libraries should process, index, and present these materials while simultaneously addressing preservation and authenticity concerns. Such questions include: What is the research value of such an archive? How important is the physical artifact? Do researchers need access to exact systems emulation? Is providing search and browse access to the data sufficient or will researchers be interested in Rushdie's original directory structure? Once data is migrated from the original environments, do we continue to maintain those outdated systems? How do archives sustain both master and access instances of born-digital archives?

As digital librarians and archivists at Emory begin processing the born-digital components of this important archive, they must keep these questions, and the host of secondary concerns circulating around them, in the foreground of workflow and process discussions. In this paper, I will argue for the importance of balancing the urgent needs of data and system stabilization with the more long-term challenges of considering the ideal outcomes and products of processing and providing access to a rare and unique born-digital archive. This talk will track the early stages of processing the physical and digital materials comprising Rushdie's digital archive, outline approaches to handling the more complex processing requirements, discuss proposed approaches to presenting the archive to both local and distant researchers, and generalize observations drawn from the experiences with this born-digital archive to broader implications within digital libraries, digital curation, and humanities computing.

With the acquisition of any archive, a research library takes on multiple responsibilities to preserve, index, and provide access to the rare and unique materials. Libraries—digital, brick and mortar, or otherwise—must "incorporate the component of stewardship over a collection."[2] Such stewardship carries with it important responsibilities, especially in cases of archival materials. Thus, with the arrival of the first shipment of Rushdie's digital archive, consisting of three out of the five computers, our library and archive staff faced both immediate preservation demands and distant challenges for archival curation. We elected to produce a workflow that is both staged and modular, which I will only summarize here. The first task is to provide a secure and stable environment for the machines themselves. As our archives had not previously included technological artifacts, this first step required some intensive space and environment analysis. Once we stabilized the physical objects, this born digital archive next challenged staff with questions of data recovery, data preservation, and data duplication. Such challenges prompted us to develop a partitioned data architecture that duplicates and preserves all master data. The original is preserved, a master duplicate is generated and stored darkly, a duplicate collection of the master database is housed in a secure repository for in-house processing and staging, and, finally, a fully-processed instance of the data is made available through a production database. Such an approach provides for preservation of original artifacts and master data, while ensuring a level of security for data while it is being processed for embargoed material.

In addition to preservation and security, authenticity of the archived data is of particular importance to archivists, digital librarians, and humanities researchers. Graham Barwell's discussions about originality and authenticity within the fields of textual studies and electronic textuality resonate with archives such as Rushdie's born-digital materials.[3] How can we most authentically represent the digital archives included in the Rushdie collection? Is the data the only component of real research value or is the context that holds this data, the paratextual elements, if you will, of equal importance to researchers? I will explore these questions and provide illustrations from our processing of the Rushdie archive to offer some preliminary insights.

---

1. From Besser's essay, "The Past, Present, and Future of Digital Libraries" in *A Companion to Digital Humanities*, 2004, p. 557.
2. Besser, p. 559.
3. Graham Barwell, "Original, Authentic, Copy: Conceptual Issues in Digital Texts" in *Literary & Linguistic Computing* 20.4 (1995) see pages 416, 418, and 419.

Markup and the Digital Paratext

Julia Flanders *(Julia_Flanders@brown.edu)*
Brown University
Domenico Fiormonte *(fiormont@uniroma3.it)*
Università Roma Tre

Abstract:

According to the French rhetorician and literary critic Gerard Genette a paratext is a transitional zone: a privileged space of a "pragmatics", of "strategy", and of action over the audience (Genette 1997). In a digital environment, the idea of the paratext may allow us to recognize the significance and the mode of operation of certain crucial textual formations that might otherwise seem inconsequential. These formations, unlike traditional paratexts, are not always literally visible as part of the legible textual surface, but instead operate through the representational and performative mechanisms of the digital interface. Their separateness is demarcated through markup, which not only creates a boundary between text and paratext but also makes the paratext into a space of function and behavior: of meaning instantiated through action rather than simply through textual signification.

Proposal:

According to the French rhetorician and literary critic Gerard Genette a paratext is a transitional zone: a privileged space of a "pragmatics", of "strategy", and of action over the audience (Genette 1997). This paratextual space—instantiated in what are traditionally considered secondary or ancillary texts such as footnotes, commentary, translation, and so forth—produces and ramifies the "main" text, the object that is thus presented to our interest as primarily significant. It is precisely through the paratexts that the concept of a "main" text emerges at all: in a given historical moment, depending on social, cultural, political, and other factors, a certain text emerges, and slowly starts to circulate and acquire status through the effects of its paratexts: comments, editions, translations, dedications, and so forth.

In a digital environment, the idea of the paratext may allow us to recognize the significance and the mode of operation of certain crucial textual formations that might otherwise seem inconsequential. An example of such a formation is the information we might call "microtexts": that is, external, contingently visible “microcontents” (Nielsen 2000) that describe, illustrate or complete information within a web page or resource. These include also embedded alternate texts that usually appear when you move the mouse over an image, link or any other semiotic device (Zinna 2005). Another example is what we might call "metatexts", a category embracing both formal metadata and also looser kinds of descriptive, normative or regulative information, usually not directly visible to the user, that is added in order to allow search engines and other processing to produce coherent and useful results.

These terms arise from an external classification (one which could be extended) emphasizing the different ways these paratexts function as part of a publication architecture. But digital paratexts can be also organized, according to their cognitive and semiotic functions, into at least three flexible (and mutually permeable) categories:

1. 1. descriptive (syntagmatic and paradigmatic axes): those paratexts which contain information about a text, including various kinds of metadata;
2. 2. normative: those which constrain the behavior of the text (for instance, schemas)
3. 3. pragmatic: those which mediate or represent the text as a discursive object, and which produce its digital phenomenology.

These formations, unlike traditional paratexts, are not always literally visible as part of the legible textual surface: their "strategy" and "pragmatics" operate through the representational and performative mechanisms of the digital interface. In print, the paratext has (despite its marginality) a certain visible presence on the page: its meaning may be wholly in relation to the text it supports, but it is not different in kind from that text (Tomasi 2005). The digital paratexts we are describing, however, occupy a different stratum and have a different kind of visibility and functional effect. Their presence may be felt primarily in what the text does or in how we discover it, rather than in the words we see when we read it. Their separateness is demarcated through markup, which not only creates a boundary between text and paratext but also makes the paratext into a space of function and behavior: of meaning instantiated through action rather than simply through textual signification. In this respect, these categories of paratext (and in particular the last, which is the realm of markup proper) are analogous to the punctuation, formatting, and presentational conventions through which a printed text is realized for a reader, or through which an oral text is concretized in print. As Genette observes, any transcription, including the written transcription of an oral speech, is a form of paratext, and we can usefully extend this idea to digital forms by observing that any encoding—in effect, any transmediation—constitutes a form of paratext as well.

Each of the three categories above illustrates a distinctive dimension of the digital paratext, and raises a set of questions and issues that can help us reach a clearer understanding of the role of markup and the nature of digital texts. If we consider markup in Genette's terms as a "privileged space" of "action" and "pragmatics", what effect does this have on our understanding of text encoding as a transcriptional activity? To what extent do our disciplinary expectations about documentary evidence run counter to the logic implied by these terms? The boundary between transcription and authoring (and our sense of where the distinction lies) may turn out to be harder to pin down than has previously been suggested.

Similarly, to what extent does metadata operate as an active, rather than a passive component of the textual ecology? If we understand metadata creation as part of the authoring of the digital document, does this change our sense of who should be creating it and what its sources should be?

Finally, if schemas and stylesheets can also be understood as paratexts (for instance, as suggested in Pierazzo 2006), then we need to rethink our conception of textual meaning and rhetoric to include not only the forms a text actually takes, but also those which its constraints permit it to take: in other words, the potential as well as actual forms through which we apprehend the text.

Bibliography

Genette, Gerard. *Paratexts: Thresholds of Interpretation.* Trans. Jane A. Lewin. Cambridge: Cambridge University Press, 1997.

Nielsen, Jakob. *Designing Web Usability: The Practice of Simplicity*. Indianapolis, IN: New Riders, 2000.

Pierazzo, Elena. "Just Different Layers? Stylesheets and Digital Edition Methodology." Paper presented at Digital Humanities 2006, Paris Sorbonne, 5-9 July 2006. 2006.

Tomasi, F. "Il paratesto nei documenti elettronici." *I dintorni del testo: approcci alle periferie del libro. Atti del Convegno internazionale. Roma, 15-17 novembre 2004; Bologna, 18-19 novembre 2004.* Ed. M. Santoro and M. G. Tavoni. Roma: Edizioni dell'Ateneo, 2005. 712-722.

Zinna, A. *Le interfacce degli oggetti di scrittura.* Roma: Meltemi, 2004.

The Voyage of the Slave Ship Sally: Exploring Historical Documents in Context

Julia Flanders *(Julia_Flanders@brown.edu)*
Women Writers Project
Brown University
Kerri Hicks *(Kerri_Hicks@Brown.edu)*
Brown University
Clifford Wulfman *(clifford_wulfman@brown.edu)*
Brown University

The Voyage of the Slave Ship Sally is a digital project developed as part of the work of Brown University's Steering Committee on Slavery and Justice, which in 2003 was formed to investigate the university's historical relationship to slavery and the slave trade. The committee's research included working with archival materials from the John Carter Brown Library and the John Hay Library at Brown University. A significant subset of these materials concerned a single slave voyage conducted by John and Nicholas Brown in the 1760s, for which a large amount of documentary evidence survives, including ship's manifests, invoices and receipts, letters, bills of lading, and the captain's trade book. The Brown University Scholarly Technology Group (STG) worked with Professor James Campbell, chair of the committee, and Patrick Yott, director of the Center for Digital Initiatives, to develop a resource through which these documents could be read and explored.

The goal of the project was to present these materials in a digital form that would respond to the committee's expressed goal to "help the campus and the nation reflect on the meaning of this history in the present, on the complex historical, political, legal, and moral questions posed by any present-day confrontation with past injustice." The project thus presented a number of challenges: the breadth of audience, the need to engage readers with widely varying levels of familiarity with the historical background, and the need to offer as many ways as possible of engaging with and probing these materials, so that they would yield the maximum possible insight into the social, political, historical, cultural, and economic circumstances surrounding the slave trade. The initial version of the site, which was released concurrently with the committee's report, emphasizes access to the primary source documents through page images

and TEI transcriptions. The second phase of the project's development, which will be completed by the end of 2007, will provide more detailed forms of access through specialized interface tools that draw on additional contextual data being developed by the project faculty.

This poster and demonstration focuses on three areas of particular significance arising from this project:

1. The interface design, which emphasized ways of engaging readers of all skill and education levels and providing them with contextual information needed to interact meaningfully with the primary source documents
2. The management of interconnections between TEI transcriptions, metadata, images, and contextual data, which support a set of exploratory tools.
3. The exploratory tools themselves, which encourage readers to probe the historical sources and their context in ways that go beyond the authoritative information provided by the site itself.

Interface design

One of the central premises guiding the project's design was the fact that we were anticipating readers in two broad categories: those familiar with the historical context who would be immediately interested in exploring the primary source documents and able to make some sense of them, and those coming to the site with no prior knowledge, who might have no initial interest in the documents at all and no way of exploring them meaningfully. Readers in the former group would need powerful tools to allow them to probe the documents, while those in the latter group would need a more narrative interface emphasizing the historical context and its modern significance, and providing readers with a sense of why the primary source documents constitute an important intellectual resource. The interface is designed with the aim of leading all readers, eventually, to the documents via links from the historical narrative to the relevant documentary material, but it also provides readers with search and exploration tools that embody as much contextual information as possible: glossed timelines and maps, social networks of people, and lists of commodities involved in the slave trade, all drawn from the information carried in the primary source documents themselves.

Data Structures

The documents and contextual information are represented through a set of interrelated data structures, specifically:

- TEI transcriptions of the primary source documents
- digital images of each document at a variety of resolutions
- METS records representing the structure of each digital object
- a set of simple XML data structures to represent four categories of contextual information on persons, places, events, and commodities named in the texts, to allow for glosses and information about the interconnections between items in these groups (such as social networks, family relationships, classifications of commodities, and so forth). The TEI transcriptions capture a comparatively large amount of information about the documents' content, by encoding all references to persons, places, commodities, and financial transactions so that these can be indexed, glossed, regularized, and searched. The transcriptions use this encoding to reference the contextual information, using key= attributes which point to the relevant records describing the person, place, event, or commodity in question. This allows both for basic hooks to allow (for instance) links to glosses directly from the text, and also for more advanced exploratory tools which process this information and represent it through visualizations and advanced search mechanisms.

Exploratory tools

The primary source documents for the Sally project are often individually compelling, but they are most significant when understood as part of a social and economic network. The existence of this network, however, can only be grasped indirectly and fragmentarily through the documents themselves; the reader needs tools for exploration that allow patterns and connections to become visible, and that can draw on the contextual information provided. When completed the project will include a set of interface tools that facilitate precisely this kind of exploration. For instance, the reader will be able to explore the network of people and social relations that surround the Sally voyage and inhabit its documentary record--owners, captain, seamen, outfitters, workmen, employees, enslaved Africans, slave traders. Their documentary relationships (for instance, the senders and receivers of letters and invoices) will be visible along with all of the metadata about those exchanges (place and date of writing, type of document, etc.) so that readers can construct a detailed and vivid picture of the vectors of communication through which this voyage is documented. Similarly detailed information supports the exploration of the commodities described in these documents, both those that were explicitly traded for slaves (rum, cloth, iron bars), and those that were structurally essential to the trade (weapons, sugar, the ship's provisions). Visualization tools such as word clouds, manipulable representations of networks, and generated timelines will give readers a variety of ways to explore this information in ways

that allow an overall grasp of the landscape as well as direct connections to the documentary evidence at every point.

Conclusions

Because the collection of documents associated with the Sally voyage is comparatively small (on the order of a hundred or so), the greatest informational value was to be gained by an emphasis on detail (through the encoding of the transcriptions) and on contextual information (through the associated XML data structures that represent ontological relationships between the various persons, commodities, and so forth). This kind of approach is uncommon for collections of historical documents, which characteristically emphasize scale (consider for instance the Valley of the Shadow, or the Making of America projects). Using visualization tools to model and interact with information of this type is also comparatively rare; visualization tools are being increasingly used for exploration of very large data sets (often not encoded in XML at all) but are less common for small, heavily tagged document collections. This project thus offers opportunities for an exploration of digital methods which may yield some useful results at its conclusion--in particular, on the question of how much value is added by detailed markup and tools that exploit it.

Round Table: Coalition of Digital Humanities Centers

Neil Fraistat *(fraistat@umd.edu)*
University of Maryland
John Unsworth *(unsworth@uiuc.edu)*
University of Illinois at Urbana-Champaign
Katherine L. Walter *(kwalter1@unl.edu)*
University of Nebraska - Lincoln
Julia Flanders *(Julia_Flanders@brown.edu)*
Brown University
Matthew Kirschenbaum *(mgk@umd.edu)*
University of Maryland

The members of the panel will discuss with the audience the outcome of the national summit meeting of American digital humanities centers and major funders held at the National Endowment of Humanities headquarters on April 12-13, co-sponsored by the NEH and the University of Maryland.

The purpose of the meeting was to take seriously the ACLS Cyberinfrastructure Commission's call for digital humanities centers to become key nodes of cyberinfrastructure in the United States. To that end, the summit explored how best to foster collaboration among national digital humanities centers, among major funders of the digital humanities, and between centers and funders, along with the possibility of creating a national coalition of digital humanities centers.

Among the salient issues are the creation of mechanisms for promoting timely exchange of information between centers and identifying opportunities for collaboration among them; the articulation of mutual research goals as well as particular areas of emphases so as to avoid duplication of effort among centers; the development of joint policy agendas and successful strategies for pursuing them; better articulating funding needs to funding agencies; methods for optimizing collaborative opportunities between humanities content specialists and the technological community; and advancing the place of the digital humanities on campus, including such issues as infrastructure and the role of the Digital Humanities in promotion and tenure.

This roundtable session will focus upon the initiatives begun at the summit meeting and engage in a general discussion about them, with the hope of widening the conversation to include

other American centers not represented at the summit meeting and centers in other countries.

Extracting Stylistic Distances from Texts for Forensic Linguistics Purposes

Katerina T. Frantzi *(frantzi@rhodes.aegean.gr)*
Department of Mediterranean Studies
University of the Aegean

The paper proposes a method for the automatic processing of forensic texts for the extraction of similarities and dissimilarities to be used for authorship identification purposes. Forensic Linguistics expresses the application of linguistics to the analysis of forensic language in written or oral form, the study of the language of the law, the study of legal interpretation and translation, the alleviation of disadvantage produced by language in legal processes, the provision of forensic linguistic evidence and of linguistic expertise in issues of legal drafting and interpretation (Conley & O'Barr, 1998; Shy 1998; Cotterill, 2002; McMenamin & Dongdoo 2002; Semino & Culpeper, 2002; Gibbons, 2003). Our forensic texts belong to a Greek terrorist organization, "17 November" (Kassimeris, 1997; Παπαχελάς & Τέλλογλου, 2002; Frantzi, 2004). 17N started its action in 1975, while the summer/autumn of 2002, all (or most) of the members were arrested. The case is currently in court for the second time. 84 of the proclamations, that used to follow a 17N attack, have been characterized as genuine (Κάκτος, 2002). Their authorship is an open issue. The final aim of this project is to map each of the 17N proclamations to its possible author (Frantzi & Ananiadou, 2006). In this work we propose a way to characterise an organization's text, i.e. proclamation, pleading, replication, by values on its stylistic characteristics. The values are used for extracting similarities and dissimilarities among texts.

We consider the following stylistics features: the average sentence and word length, the number of different nouns, proper nouns, adjectives, active voice verbs, passive voice verbs, adverbs, the number of future tense references, the use of conjunctions, preposition, pronouns, particles, determiners, articles, question words, punctuation, content words. For each pair of possibly owned by the same author writings, we create a matrix which we call *distance-matrix* (Table 1). Each cell of the matrix contains a number that corresponds to a stylistic feature distance between the two writings.

particles-distance	determiners-distance	articles-distance	conjunctions-distance	pronouns-distance
prepositions-distance	question-words-distance	future-refs-distance	proper-noun-distance	adverbs-distance
sentence-length-distance	adjectives-distance	passive-verbs-distance	nouns-distance	active-verbs-distance
content-words distance	? distance	! distance	: distance	word-length distance

Table 1. Contents of a *distance-matrix*.

In order to compute the *distance-matrix* for a pair of writings, we need to assign a value to each cell that corresponds to a particular stylistic feature. This value gives the difference (distance) between the two writings regarding the specific feature. Then the distance between the two writings is evaluated by the average value of all the *feature-distances* in the *distance-matrix*. A *feature-distance* is assigned to a cell and is calculated by the average of all phenomena distances for that feature. The phenomena distances for a feature are calculated by their normalized frequencies of occurrence in the two writings. The algorithm for the processing of the *feature-distances* is the following:

for each pair (writing$_1$, writing$_2$)

for each cell *i* in *distance-matrix*(writing$_1$, writing$_2$)

assign *feature$_i$_distance*(writing$_1$, writing$_2$)

distance(writing$_1$, writing$_2$)= avg(for all *i*, *feature$_i$_distance*(writing$_1$, writing$_2$))

feature$_i$ _distance(writing$_1$, writing$_2$)=avg(*feature$_i$ _distance*(writing$_1$, writing$_2$)[p_j])

for each p_j phenomenon of *feature$_i$*

feature$_i$_distance(writing$_1$, writing$_2$)[p_j]= $|f_{1j}-f_{2j}|/(f_{1j}+f_{2j})$

where

f_{1j} and f_{2j} are the frequencies of occurrence of the p_j phenomenon for writing$_1$ and writing$_2$

p_j is a phenomenon of *feature$_i$*

Each author is characterized by his/her pleadings and replications. If we could match a proclamation to a pleading (and replication) then we could possibly match a proclamation to the corresponding author. We need to compare all pleadings (and replications) to the proclamations, to find the smallest "distance" in terms of their feature values. The smallest distance assigns a proclamation to a pleading, i.e. an author.

In this work we apply the *distance-table* method for comparisons between the pleadings and replications: we already know the authors of the pleadings and proclamations, so we can present how the *distance-table* works. We apply the method to three writings: one pleading and two replications. Let us consider Yotopoulos's (the person accused to be the main leader and one of the main authors of the proclamations) pleading of 18321 words, and replication of 2036 words, and Koufontinas's (also accused to be a leading member) replication of 326 words. We will evaluate the *distance-matrix* for the pairs of writings (Yotopoulos's pleading, Yotopoulos's replication) and (Yotopoulos's pleading, Koufontinas's replication). We will show the evaluation for one of the cells of the two *distance-matrixes*, the *distance-matrix*[1,1], i.e. the cell that keeps the distance regarding the stylistic feature of the particles' use, while the evaluation is analogous for the rest of the cells.[1] There are five particles in Greek: "να", "για", "θα", "ας" and "μα". Table 2 gives the frequencies of occurrence and normalized frequencies (to 10,000 words) for the five particles found in Yotopoulos's pleading (Yp) and replication (Yr) and those in Koufodinas's replication (Kr).

Y_p			Y_r			K_r		
να	450	245.619	να	60	294.695	για	6	184.049
για	183	99.885	θα	22	108.055	να	4	122.699
θα	165	90.060	για	20	98.232	θα	3	92.024
ας	6	3.274	να	0	0		0	0

Table 2. The particles' frequencies of occurrence and normalised frequencies for the 3 writings.

The *distance-matrix*[1,1] cell for the (Yotopoulos's pleading, Yotopoulos's replication) *distance-matrix* holds the distance regarding the use of particles. We evaluate the distance on the use of each of the five particle between the two writings and we assign to the *distance-matrix*1,1] cell the average of all the particles' distances. Then we do the same for the *distance-matrix*1,1] cell of the (Yotopoulos's pleading, Koufodinas's replication) *distance-matrix*. The particles' distances for the particle "να" for the two pair of writings are calculated according the above-given algorithm as:The *distance-matrix*[1,1] cell for the (Yotopoulos's pleading, Yotopoulos's replication) *distance-matrix* holds the distance regarding the use of particles. We evaluate the distance on the use of each of the five particle between the two writings and we assign to the *distance-matrix*[1,1] cell the average of all the particles' distances. Then we do the same for the *distance-matrix*[1,1] cell of the (Yotopoulos's pleading, Koufodinas's replication) *distance-matrix*. The particles' distances for the particle "να" for the two pair of writings are calculated according the above-given algorithm as:

Particles_distance(Yp,Yr)[να]=|245.619 - 294.695| / (245.619 + 294.695)=0.091 Particles_distance(Yp,Kr)[να]=|245.619 - 122.699| / (245.619 + 122.699)=0.334

Table 3 gives the results of the calculations for all the particles' distances for those pair of writings. We get the average for each of the columns of Table 3, i.e. 0.289 and 0.409. These averages will be assigned to *distance-matrixes.*[1,1] cells for the two pairs of writings *distance-matrixes.*

	Particles_distance(Y_p,Y_r)	Particles_distance(Y_p,K_r)
να	0.091	0.334
για	0.0088	0.251
θα	0.0587	0.051
ας	1	1
avg	0.289	0.409

Table 3. The particles' distances for the pairs of writings (Y_p,Y_r) and (Y_p,K_r).

Table 4 gives all the *feature-distances* for the (Yotopoulos's pleading, Yotopoulos's replication) *distance-matrix*while Table 5 for the (Yotopoulos's pleading, Koufodinas's replication) one.

0.289	0.225	0.191	0.552	0.63
0.32	0.225	0.09	0.305	0.34
0.232	0.04	0.17	0.017	0.009
0.616	0.477	1	1	0.036

Table 4. The *distance-matrix* for the pair of writings (Y_p, Y_r).

0.409	0.923	0.667	0.922	0.849
0.622	0.776	0.010	0.575	0.474
0.176	0.27	0.115	0.485	0.073
0.798	1	1	1	0.042

Table 5. The *distance-matrix* for the pair of writings (Y_p, K_r).

When matrixes are filled up, we can evaluate the total distances for the pairs of writings as the average of all the cell values:

matrix-distance(Yp, Kr)= 0.5593 *matrix-distance* (Yp,Yr)= 0.3382
As a result Yotopoulos's pleading is linked to Yotopoulos's replication (smaller distance), which actually means that these two writings have been found to be closer as for their language style than the writings of Yotopoulos's pleading and Koufodinas's replication.

Future work involves:

- Augmentation of the *distance-matrix* with more stylistics features, e.g. the use of collocations (Frantzi & Ananiadou, 2007),
- use of weights for the stylistics features,
- application of the method for characterizing the proclamations for the provision of authorship evidence,
- comparisons among the proclamations for their grouping according to their linguistic profile.

1. We cannot present the evaluation of all cells here due to space restrictions.

Bibliography

Conley, John M., and William M. O'Barry. *Just Words – Law, Language, and Power*. Chicago and London: The University of Chicago Press, 1998.

Cotterill, Janet, ed. *Language in the Legal Process.* MacMillan, 2002.

Frantzi, Katerina T. "Corpus Linguistics: What can it do with Terrorism?" *International Journal of Humanities* 2.2 (2004): 1603-1608.

Frantzi, Katerina T., and Sophia Ananiadou. "Automatic Authorship Identification." *Proceedings of the International Association of Forensic Linguistics 2nd European Conference on Forensic Linguistics / Language & the Law*. Barcelona, Spain. 14-16.

Frantzi, Katerina T., and Sophia Ananiadou. "C-value for Authorship Identification." *Proceedings of the International Association of Forensic Linguistics 8th Biennial International Conference on Forensic Linguistics / Language & the Law – IAFL 8, Seattle, Washington, July 12-15 2007.* forthcoming.

Gibbons, John. *Forensic Linguistics: An Introduction to Language in the Justice System.* Language in Society 32. Blackwell Publishing Ltd, 2003.

Kassimeris, George. *Europe's Last Red Terrorists: The Revolutionary Organization 17 November*. C. Hurst & Co. Ltd, 1997.

McMenamin, Gerald R., and Choi Dongdoo, eds. *Forensic Linguistics : Advances in Forensic Stylistics.* CRC Press, 2002.

Semino, Elena, and Jonathan Culpeper, eds. *Cognitive Stylistics: Language and Cognition in Text Analysis.* Linguistic Approaches to Literature 1. John Benjamins Publishing Company, 2002.

Shuy, Roger W.>. *The Language of Confession, Interrogation and Deception* . Empirical Linguistics Series. Sage Publications, 1998.

Κάκτος. *17Ν – Οι προκηρύξεις.* Αθήνα: Κάκτος, 2002.

Παπαχελάς, Α., and Τ. Τέλλογλου. Αθήνα, Ελλάδα: Εστία, 2002.

Ancient Technical Manuscripts: the Case of 17th-century Portuguese Shipbuilding Treatises

Richard Furuta *(furuta@cs.tamu.edu)*
Center for the Study of Digital Libraries
Texas A&M University
Filipe Castro *(fvcastro@tamu.edu)*
Center for Maritime Archaeology and Conservation
Texas A&M University
Carlos Monroy *(cmonroy@cs.tamu.edu)*
Center for the Study of Digital Libraries
Texas A&M University

During the 16th- and 17th-centuries European seafaring underwent an incredible transformation driven mainly by the exploration of newly discovered lands, the contact with previously unknown cultures, and the increase in maritime commerce. The social changes that resulted from this cultural revolution affected the long chain of events entailed by the construction of oceangoing ships and determined a number of technical innovations in the construction of ships. From an original oral tradition, where apprentices learned from masters the intricacies of shipbuilding techniques; this process evolved into a more formal field as masters began to follow guidelines, materials used, and construction sequences in a more systematic way, forming a corpus of information that was soon compiled in manuscripts known as shipbuilding treatises.

The first shipbuilding treatises, understandably, were written by mathematicians, priests and other learned men, reflecting a reality where shipbuilders probably were largely illiterate. Rich in technical descriptions, shipbuilding treatises play a key role in Nautical Archaeology both for scholars and students. Scholars access these manuscripts for several reasons. For example, working on the reconstruction of sunken ships, they can provide vital information for reassembling fragments and damaged timbers of ship remains recovered from underwater excavations. Also, their contents often help our understanding of shipbuilding techniques. In addition, they can be used to compare different construction traditions both from geographical and chronological standpoints. Moreover, they are great sources for understanding the evolution of shipbuilding (Figure 2 shows our treatises browser).

Nautical Archaeology students—although not carrying out the complex tasks of ship reconstruction—are exposed to treatises early in their studies. For them treatises are a good source to understand the basic terminology and concepts they will be using during the rest of their studies. Current teaching practices are constrained to browse physical copies of the original sources; with obvious restrictions such as limited number of copies and access, or unknown language—depending of their provenance, they were written in different languages, and difficult technical terminology. In fact, treatises inherit most of the limitations of printed books.

Providing ways in which shipbuilding treatises can be used digitally by both scholars and students is an attractive interdisciplinary effort for a number of reasons. First, is the opportunity to investigate the variety of ways in which manuscripts' contents (both texts and illustrations) can be structured and classified. Second, is the opportunity to make original-source material available at the location of an excavation; the treatises description of the construction of the physical objects can provide valuable information about fragments of ship timbers that are recovered from an excavation site. This effort clearly draws techniques from the earlier projects in the digital humanities that examine the characteristics of digital representations of paper-based texts. However the linkage to physical artifacts opens up additional possibilities and considerations.

Digital humanities projects involving manuscripts or printed texts have been related, for the most part, to literature and historical records; some Know: Responding to the Computational Transformation of the well-known examples include the Canterbury Tales Project.[1], the Rossetti Archive[2], and the Perseus Digital Library [3]. We have been involved in creating collections of this form as well for Cervantes[4],Donne,[5] and Picasso[6]. In the context of Nautical Archaeology, the manuscript of Michael of Rhodes captures the knowledge obtained by this 15th century seaman during his 4- decade-long career[7]. The presentation of the manuscript, oriented to a general audience, hints at the value that such materials will have to the professional archaeologist when representations and tools are provided that meet his scholarly needs.

Until relatively recently, ships were the most advanced and complex transportation means designed. Nautical treatises hold the key to understanding their technical complexity. In essence, the collection of treatises represents the technical manuals describing the components, their use, and the steps taken in manufacturing of the ship. Several characteristics of treatises make understanding them a very challenging task. Language is a major problem; in order to better understand their contents, it is necessary to provide translations and explanations of

concepts, pieces, and sequences. Beyond the multiple languages in which the treatises were written, they also come from diverse geographical locations and span centuries, making terms, concepts, and descriptions difficult to understand. Different units and standards of measurement—a key aspect in technical descriptions—raise problems about not only comparing treatises with different provenance, but also translating them into modern scales; units of measurement used in the treatises are not necessarily the ones used by archaeologists to measure recovered evidence. To tackle these problems, we have developed a multilingual glossary, in which terms include their corresponding translation and definition into ten languages (which can be expanded as needed). The incorporation of "roles," enables us to expand characteristics related to the terms, for example spellings and synonyms. Our framework allows multiple values per role as well the addition of more roles as they are required (see figure 1).

However, despite common features shared by physical fragments and their corresponding descriptions in the texts, fragments obtained from individual ships have important differences because of the differing physical conditions that they have been exposed to. Damaged and incomplete ship remains require the adoption of an encoding scheme to describe and quantify uncertainty; textual descriptions do not encompass "uncertainty."

In the previous paragraphs we have briefly outlined the relationship between physical archaeological evidence and written descriptions in the treatises. However, treatises in themselves have properties that make them unique. For example, an important question is how similar or different are treatises in terms of the sequences, construction techniques, and materials used. An initial approach would suggest that probably the encoding used in their description could be used to quantify the degree of similarity.

Since treatises are "technical manuals," illustrations are essential in their understanding; therefore, we adopted a two-step process. First, illustrations have to be segmented, an illustration *Ij*, can be composed of a set of components $C = \{c1, c2,\ldots ck\}$, where a component *ci* has a list of properties $P = \{p1, p2 \ldots pk\}$. Second, each component might have a description within the text, thus a linkage between the two is required. To make things more complex, a component *cj* can be formed from a subset of components, a step that resembles a recursive property, where the ship as a whole is formed by small parts, which in turn are composed by smaller ones, and so on. Figure 3 depicts the interface for capturing coordinates in images linking them to terms from the glossary.

Conversely, components can be mapped to other representations; a good example is a model created in 3D rendering software such as Rhino. We have done preliminary tests, exporting geometric data from Rhino models into XML and linking them to both 2D slides of the model and their corresponding occurrences in the treatises.

Although linking text and images has been extensively studied; the context of treatises raises a series of complex issues. For example, the text of a treatise could be segmented in a variety of ways based on different needs; assemblage sequences, materials used, and section of the ship being described. This in turn raises some interesting questions, for example: could the components being included in part of the text give a hint of what that section is about, or what section of the ship it describes? How could components in different treatises be compared?

Our current collection includes digital images and transcriptions of three of the most relevant late 16th and early 17th century Portuguese treatises: [8] Fernando Oliveira's *Livro da Fabrica das Naos* (dated to 1580), João Baptista Lavahna's *Livro Primeiro da Architectura Naval*[9] (dated between 1608 and 1615), and Manoel Fernandez's *Livro de Tracas de Carpintaria* (dated to 1616). We expect to add more manuscripts as permissions from holders are granted.

The treatises' dual role as historically-significant text and as formal specification of elements of ship design affords an opportunity to investigate the relationship of techniques developed within the context of textual studies to applications with physical objects and virtual 3D models. Further, the treatises provide the foundation for our development of the Nautical Archaeology Digital Library (<http://nadl.tamu.edu/>), which will center on providing resources in support of archaeologists' work and on dissemination of expedition artifacts[10]. The combination holds promise of extending the reach of the digital humanities.

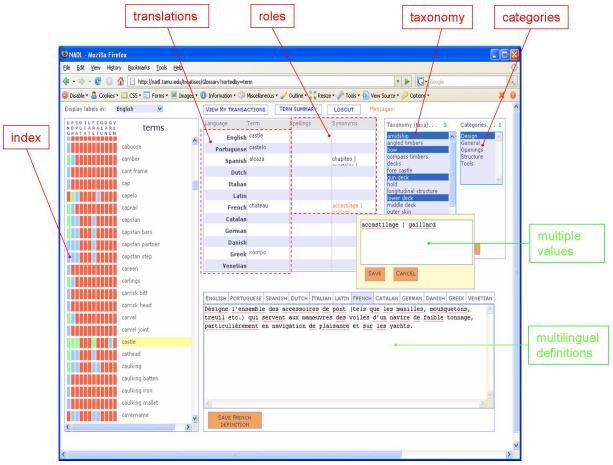


Figure 1: The multilingual glossary interface depicting terms, translations, and definitions in various languages.

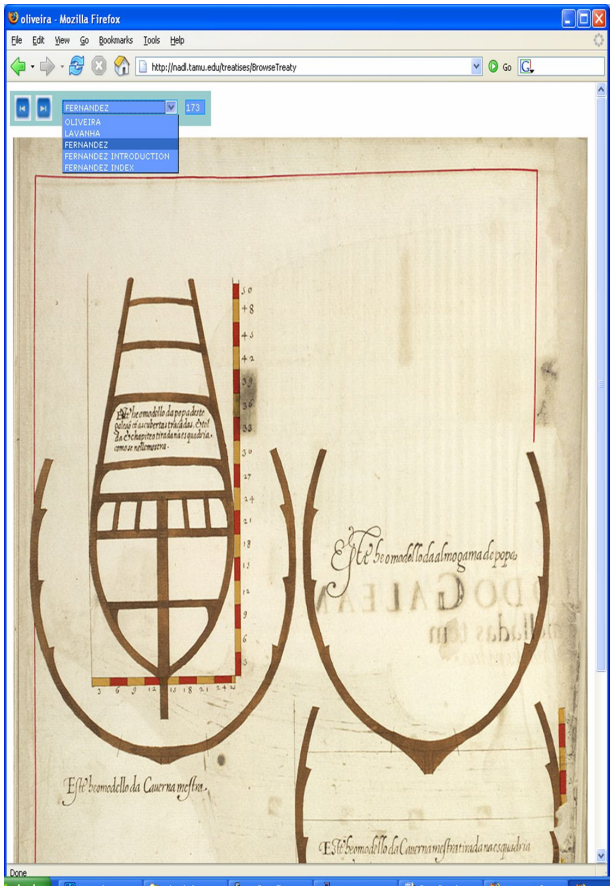


Figure 2: Treatises browser allows navigation of the treatises.

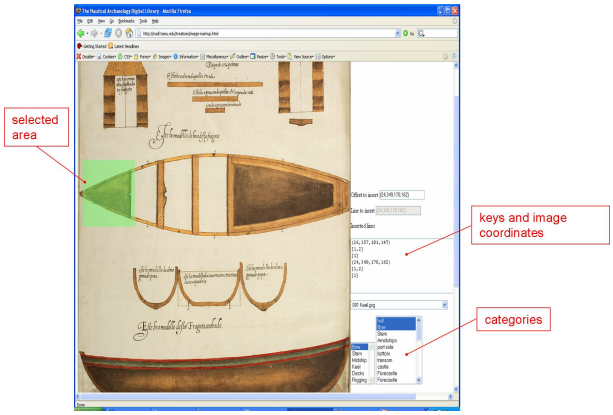


Figure 3: Treatises markup utility allows to mark certain areas of the image and associate them with terms and categories in the glossary.

1. The Canterbury Tales Project Accessed 2006-11-12. <http://www.canterburytalesproject.org/>
2. The Rossetti Archive Accessed 2006-11-12. <http://www.iath.virginia.edu/rossetti/>
3. The Perseus Digital Library Accessed 2006-11-12. <http://www.perseus.tufts.edu/>
4. Eduardo Urbina, Electronic Variorum Edition of the *Quixote* Accessed 2006-11-12. <http://www.iath.virginia.edu/rossetti/>
5. Gary Stringer, Digital Donne. Accessed 2006-11-12. <http://www.csdl.tamu.edu/donne>
6. Enrique Mallen, The On-line Picasso Project Accessed 2006-11-12. <http://picasso.tamu.edu/>
7. Michael of Rhodes Accessed 2006-11-12. <http://www.iath.virginia.edu/rossetti/>
8. Special thanks to Academia de Marinha, Lisbon Portugal for providing the facsimiles. Accessed 2006-11-12. <http://www.marinha.pt/Marinha/PT/Menu/DescobrirMarinha/Actividade/AreaCultural/academia/>
9. Lavahna, Joao Baptista. *Livro Primeiro da Architectura Naval*, c. 1610. Facsimile, transcription, and translation into English, Lisbon, Academia de Marinha, 1996.
10. Dissemination of archaeological artifacts also has been the focus of other significant efforts including ArchaeoML, used in OCHRE Accessed 2006-11-12. <http://ochre.lib.uchicago.edu/>formerly XSTAR, and ETANA Accessed 2006-11-12. <http://www.etana.org/>

Digitization and Publication of the *Goethe-Dictionary* on the Internet

Kurt Gärtner *(gaertnek@staff.uni-marburg.de)*
University of Trier, Germany
Vera Hildenbrandt *(hildenbr@uni-trier.de)*
University of Trier, Germany

The Goethe Dictionary(*Goethe Wörterbuch* = GWb) is one of the most ambitious projects in German lexicography. Its importance for literary and linguistic research on the writings of the most famous German author Johann Wolfgang von Goethe (1749-1832) can not be overestimated. Although still a work in progress – hitherto four volumes (*A—inhaftieren*) have been completed – the GWb has become an essential tool not only for the study of the wide ranging vocabulary of an individual author, but also – due to the enormous influence of this author – an indispensable instrument for research on the history of the German language and culture during one of the key periods of their development. The vocabulary of Goethe comprises about 90.000 words, more than any other writer in German has ever used, because he not only wrote as a poet, but was also engaged in serious research in fields like anatomy, geology and botany; and in addition to that he served as a minister to the court of Weimar and was well familiar with all aspects of the administrative language of his time.

The preparation of the GWb began shortly after the 2nd world war in 1947, the 1st volume appeared in 1978, the second instalment of the 5th volume (*Jammernachbar—kanonieren*) in 2005. In 2003 plans were developed for the creation of a digital version of the GWb which so far had been published only in a printed form. A grant proposal for the digitization had been drawn up by the *Competence Centre for Electronic Retrieval and Publication Methods in the Humanities* (*Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren*) at the University of Trier in collaboration with the *Interacademic Committee for the Goethe Dictionary* of three German Academies of Sciences and Humanities (Berlin-Brandenburg, Göttingen, Heidelberg).

After receiving a grant from the *German Research Association (Deutsche Forschungsgemeinschaft)* the digitization project started in May 2004 at the University of Trier. By end of 2005 the first 3 volumes were made freely available on the internet (<http://www.gwb.uni-trier.de>) in a preliminary version which has been refined in the following time considerably. In future the GWb will continue to be published in print, but will be also made freely available on the Internet in a version which is closely related to the printed form with its stable references for reliable quotations. Conservative Goethe-scholars with their prejudices towards the new media may quote the book, although their secretaries got the quotations from the digital version. It is therefore essential that the digitized version is an absolutely correct rendering of the printed version and offers the same degree of reliability. However, the Internet version will offer much more than the book with its restriction caused by the procrustean bed of its alphabetical macrostructure: Via a comfortable and user-friendly interface, the electronic GWb is to provide flexible and innovative possibilities for literary and linguistic research on Goethe and the history of the German language and its vocabulary as well as for everyone interested in Goethe and the language of his time, both in Germany and abroad. The following short description of the digitization project will outline the main features of its innovative approach to making the informational potential of a traditional dictionary fully available. The various steps of the project will be described in detail in the proposed paper.

The first task in the project's working plan consisted of a minute analysis of all the typographical features of the microstructure of an entry in order to draw up a carefully organised set of instructions or functional specifications *(Pflichtenheft)* for the typists who should be enabled to apply appropriate tags related to the specific typographical functions. This preliminary tagging is done by the typist during his keying in the text of an entry; it saves a considerable amount of time in a later stage of the project when XML/TEI conformant markup is applied.

The text of the GWb has been made machine readable via double keying, a method strongly recommended by Wilhelm Ott and successfully practised in previous lexicographical projects of the Trier *Competence Centre*, e.g. the *Middle High German Dictionaries Interlinked* (*Mittelhochdeutsche Wörterbücher im Verbund*), <http://www.MWV.uni-trier.de>, and the digitization of the 33 vols. of the *Deutsches Wörterbuch* by Jacob and Wilhelm Grimm, <http://www.DWB.uni-trier.de>, the German equivalent to the OED. The idea of scanning the GWb and using OCR software for processing the data had been excluded from the beginning, because of the highly specialised typography and the far too time-consuming and costly correction processes afterwards. The data input took place in China. Two teams were capturing the GWb in MS-Word via double keying and using TUSTEP-tags for encoding the typographical features and the special characters. After the transfer of the data to Germany via e-mail, the data were converted into the TUSTEP-format. The two versions were collated automatically by TUSTEP-routines; TUSTEP also generated a list of differences, which has been corrected manually after comparing the printed

GWb. The result of this step was an error-free TUSTEP-version of the GWb.

Apart from fulltext retrieval, the TUSTEP-version already offers a number of other searches, but it is difficult to perform combined queries without further markup. For this step the TEI-Guidelines were used; they were not followed always strictly, some deviations were necessary to suit the specialities of the microstructure of a GWb-entry. Because of the huge amount of data, the XML/TEI-conformant markup had to be inserted by using automatic routines wherever possible. For this we used again TUSTEP-modules which proved to work efficiently and with an outmost degree of accuracy.

The XML/TEI-encoded data were stored in a data base and a graphical user interface was developed using several tools and methods which we had used for the previous lexicographical projects and which have been described elsewhere in detail by Thomas Burch (<http://aspn.activestate.com/ASPN/Tcl/TclConferencePapers2002/Tcl2002papers/>).

The contribution will present a detailed introduction to the user interface and the functionalities of the electronic GWb. We took great pains to display the GWb in a way that makes it attractive and easy to use. Unlike other digital dictionaries (OED, TLF) we did not allow to break up the macrostructure in single entries separately shown (one screen = one entry), but kept the structure of the printed dictionary thus allowing a comfortable overview over series of compounds beginning with the same determinative. The framing of the window allows an easy navigation especially through long entries with an elaborated hierarchical microstructure, thus helping the user to find as quickly as possible what he is looking for. – As the project is still a work in progress, the presenting author would gladly discuss the perspectives of what remains to be done.

Up-To-Date Means of Access to Full-Text Databases

Roman M. Gnutikov *(romashka@uni.udm.ru)*
Urdmurtia State University
Victor A. Baranov *(baranov@udm.ru)*
Izhevsk State Technical University

Introduction

Among mankind's priceless treasures are not only handmade artefacts but also creations reflecting mankind's ability of thinking and speaking. These are texts. Thousands, tens of thousands, hundreds of thousands of manuscripts created during the past thousands years that reached our days are of great historic, artistic and cultural value. There is no doubt that written texts were and remain one of the most valuable witnesses of the past of mankind, its achievements and discoveries, world view, sufferings and errors.

It is well known that natural human language and texts in natural language, accordingly, are among the most complicated things for study and interpretation. The complexity increases several times for old texts.

One of the most effective means of preserving and increasing knowledge of old texts is the steady philological, historical and textual study of and commentary on them using the manuscripts. Unfortunately, the major part of the world manuscript heritage is in need of the comprehensive study. Many manuscripts are unpublished and inaccessible.

Objectives

Up-to-date computer technologies offer wide possibilities for preservation, processing, study and popularisation of the manuscript cultural heritage of mankind. The most promising method of comprehensive study of manuscripts is the creation of digital libraries in the form of full-text databases. The advantages of this type of collection are the volume of stored information, speedy search, advanced data retrieval and ordering functionality, the ability to supplement the database with new information and the ability to access it through the Internet

The weaknesses of full-text databases are their complexity and the expenditure of labour required to access the manuscript/text and its components. The depth of fragmentation, number and composition of units, their properties and values all must be determined beforehand, so the text is considerably simplified and the means of establishing relationships between fragments can be absent.

While developing the Manuscript system [1] (<http://manuscripts.ru/index_en.html>) intended for storage, processing and publication of ancient manuscripts, the research group of philologists and programmers at Udmurtia State University developed a unique module for accessing the full-text database: a specialised text editor called OldEd. Its distinctive feature is its combination of the functions of traditional text processors with their visual presentation of text and formatting capabilities, with the ability to build a corresponding object-oriented database for each textual unit, which can be combined into relationship hierarchies.

As is well known, text processors like Microsoft Word are well suited to extension using their powerful object model and therefore provide a means of interaction with databases as an interface to it (in this case, with full-text databases of ancient texts). However an analysis of such a means of interacting with the database using Microsoft Word showed that the development of our own tools for editing Old Russian texts stored in the database would be considerably less labour-consuming than customizing Microsoft Word or other software

Methodology

To work with data stored in full-text databases, it is advisable to create special editors. In the process of developing the OldEd editor, our research group was guided by the following requirements:

1. creation, input and editing of the document through direct interaction with the database;
2. representation of the manuscript text in a form that would be close to the original, including reproduction of glyph variants (Figure 1);
3. selection and creation of the manuscript/text units and manipulation of their properties and values (Figure 2);
4. work with unit relationships (creation, change of subordination, deletion, change of properties, visualization of relationships);
5. work with various hierarchical structures (unit selection, creation of the relationship with the parent unit etc.) and representation of unit relationships in the hierarchy as a tree (Figure 3);
6. work with unit dictionaries, their properties and values and relationships (Figure 4);
7. support for simultaneous work with several manuscripts;
8. multi-user support.

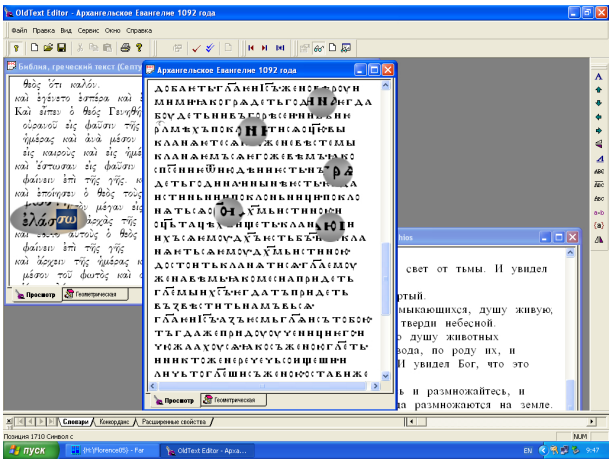


Figure 1: : Many texts, many languages and accurate reproduction of glyphs

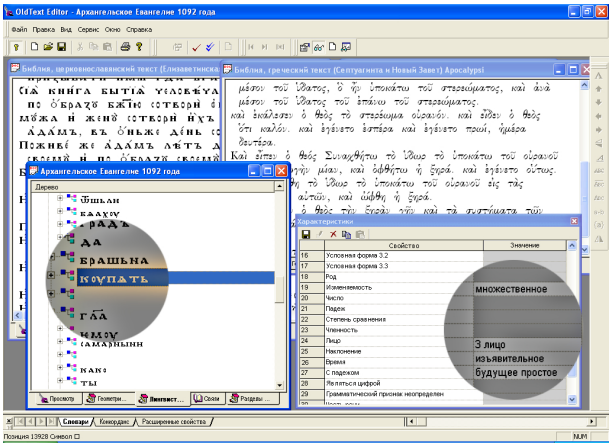


Figure 2: Text units, their properties and values

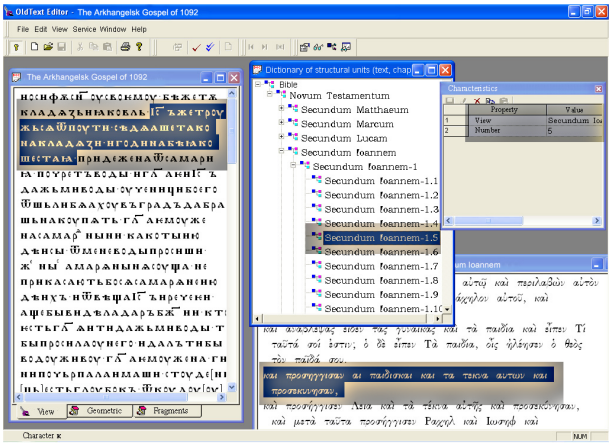


Figure 3: Text fragments and their dictionary equivalents

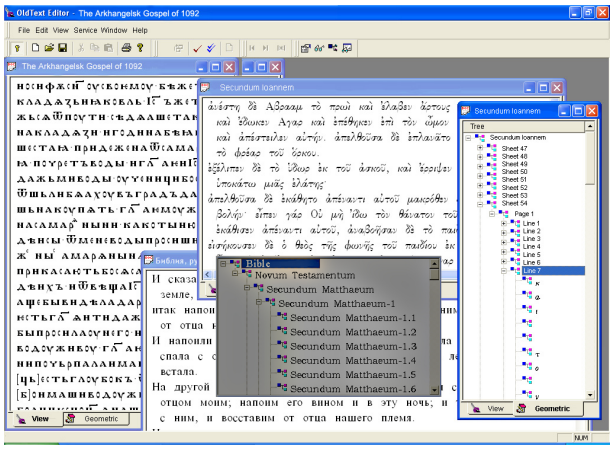


Figure 4: Hierarchic representation of units

Technological description

From a technical point of view, the editor comprises a range of components providing access to data, the ability to read and record objects into the database, visualisation of the units of the manuscripts/texts and their hierarchical relationships, and representation of the document in the form of formatted text.

The editor was written in C++, and the components were developed with the use of ATL (Active Template Library). The client part was written using MFC (Microsoft Foundation Classes). In operation the client interacts with the database of the Manuscript system by means of a server API-procedure written as packages for the Oracle database management system.

Results and Business benefits

The smallest unit that can be operated on by the editor is the glyph and its variants. The largest units are the manuscript and text. It should be noted that the relationships among the latter could be described by the notion 'many to many': the manuscript can have many texts; one text can be represented in many manuscripts.

The OldEd editor allows the user to work effectively with the visualised data – units, their relationships, properties and values – of texts/manuscripts. The editor enables representation of the hierarchies existing in a text/manuscript and of the text as a transformed geometrical hierarchy. When working with the text, the editor allows the user to edit the text and divide it into fragments. When scrolling, the editor displays information on the relationships between the text units, structural units and dictionary units; allows creation and deletion of relationships between units; allows viewing and correction of their properties and allows creation of new units (including texts). When displaying relationships, the user can view and edit all unit relationships.

Further steps

The editor's functional possibilities are being expanded in several directions:

- Since working with the editor currently requires a constant connection with the database, it would be desirable to have a feature allowing the user to download part of the document and work with it offline.
- The present version of the editor could be used with a direct remote connection to the database, but this is not allowed only for security reasons. This is why the database management system is currently behind a firewall and inaccessible from outside of the local network. This situation should be changed, as the value of the system considerably increases if many researchers can access it not only to acquire information, but also to enrich the database. To achieve this goal, a new additional web interface to data using the SOAP protocol should be developed. Information on the ancient hand-written treasures stored in the database is already accessible through a special web-interface in viewing mode (as an illustration see our site devoted to the Putyata's Menaion manuscript <http://manuscripts.ru/mns/portal.main?p1=19&p_lid=2&p_sid=1>).

Conclusions

The editor provides features for working with full-text databases: the text/manuscript and their units are represented in a view that is close to the original; the text/manuscript can be divided into fragments described by their respective properties and values; units belonging to the same content area can be organized in a hierarchy; any unit can have a standard variant (corresponding dictionary unit); relationships can be established between any units (even distant ones); and editing of units, their properties, values and relationships can be done by the user directly in the text. In other words, the digital copy of the manuscript/text can be represented not like a linear chain of units, but like a complex network, each part of which is a model of a certain area of information in the manuscript. All editor features are intended, first of all, so that versatile operations can be performed on manuscripts/texts for further processing of them in the databases, preparation of reference materials and creation of printed and digital editions.

One of the most important advantages of the editor is that it allows simple, consistent visual manipulation of the various units of structurally complex manuscripts and texts so that users do not need to learn difficult markup languages

For more detail about the editor see <http://manuscripts.ru/pub/rd/>.

Acknowledgment

The work was made possible thanks to the financial support of the Russian Foundation for Basic Research (Grant 05-07-90217-в).

1. Baranov, V.A., Votintsev, A.A., Gnutikov, R.M., Mironov, A.N., Oshchepkov, S.V., and Romanenko, V.A., "Old Slavic Manuscript Heritage: Electronic Publications and Full-Text Databases" EVA 2004 London (Electronic Imaging, the Visual Arts Conference & Beyond). Conference Proceedings. University College London, Institute of Archaeology. Principal Editor: James Hemsley. London, 2004.
Baranov, V.A., Votintsev, A.A., Gnutikov, R.M., Zuga, O.V., Mironov, A.N., Nikiforova, S.A., Oshchepkov, S.V. Romanenko, V.A., and Ryabova, E.V. (2003) "Electronnyje izdanija drevnikh pis'mennykh pamjatnikov i tekhnologija sozdanija polnotekstovykh baz dannykh (Electronic Editions of Old Manuscripts and Technology of Creation of Full-Text Databases)." Krug idej: electronnye resursy istoricheskoj informatiki, Moscow, pp. 234–260.
Baranov, V.A., Votintsev, A.A., Gnutikov, R.M., Mironov, A.N., and Romanenko, V.A., (2003) "Spetsializirivannyj tekstovyj redactor "Manuscript" Sistemy obrabotki drevnikh rukopisej (Specialized Text Editor Manuscript of the System for Processing Old Manuscripts)." Informatsionnyj bjulleten' assotsiatsii "Istorija i komp'juter 31: 159-165.

Geographical Information Systems and the Exploration of French Culture and Society

Joel Goldfield *(jgoldfield@mail.fairfield.edu)*
Fairfield University

This presentation describes the principal humanities application of a recent grant project that used Geographical Information Systems (GIS). It explores the institutional context in which the project took place, curricular and survey results, implications for faculty development using a GIS tool in the humanities, and a glimpse at where the project is leading in the study of language, literature and culture. A GIS methodology we used allowed us to map statistical information to make visible what might otherwise have been merely a matrix of numeric data. It allowed us to distinguish meaningful patterns. We have found that when the data come to bear upon historical, political or sociological situations, they can have an impact upon the study of language, culture and even literature, thus the humanities. Such was the premise of our project, the International Studies/Language Technology Initiative, funded by three American philanthropic foundations from 1999-2002, the Culpeper Foundation, the Archbold Charitable Trust, and the Rockefeller Brothers Fund.[1]

Our GIS initiative focused on International Studies, Sociology, and Modern Languages and Literatures. We sought to improve our own analytical skills across disciplinary divisions as a model for our students and to promote foreign languages across the curriculum (FLAC).[2] For French and Spanish, we introduced maps into the curriculum, combined maps with data sets, and encouraged students to answer questions in the foreign language using these materials, which involved relatively simple but relevant statistics that usually had an historical or a sociological context.

The first level consisted of faculty development, followed by the creation and pilot application of interactive maps to help students fill significant gaps in their knowledge of geography, where they could work in pairs or small groups to discuss clues about a country's characteristics or location, then click on the area within the borders to reveal the name and see if they were correct. This approach eventually led to the creation of multimedia GIS maps where the name of the country, capital city or other regions would be pronounced and where other multimedia materials such as digital video with subtitles or captioning would appear to assist language learners (Figures 1

and 2, originally in color).[3] The multimedia tagging of maps using a proprietary GIS tool, *MapInfo*, allowed us to experiment with a kind of hypertext and hypermedia suggested by Vannevar Bush's article "As We May Think" (Atlantic Monthly, July 1945, 101-108) and which have been explored by countless other researchers for a host of applications. We are currently investigating the viability of non- proprietary GIS tools such as QGIS and Google Earth applications.

In our project, we created some maps that showed change in demographic or other features over time, others that showed local, regional, national, and transnational patterns in trade, use of resources, socio-demographics and politics. Several of these custom-designed GIS maps will be shown at the presentation. Faculty anecdotally concluded that this work enhanced students' ability to read tables, recognize spatial data, read maps, and use the computer as an analytical tool, not just a word processor, e-mailing device, or means to browse the Internet.

When applied to studies and classes in modern languages and literatures, we developed pedagogies to encourage analytical and critical thinking. One successful application was to pair this author's French Translation class (FR 265) with a class in Urban/Suburban Sociology (SO 163) on several occasions, where the latter's students were charged with a task that is far from trivial, one that Martyn Jessop in the 2006 ADHO conference identified: "The number of digital datasets is growing rapidly and these are often of interest to researchers in fields other than the often highly specialised one that the data was originally derived for but how does one locate them?" (Digital Humanities 2006, p. 101). The task in our case was a "mapquest" activity to find French census data that would allow both the French Translation and anglophone Sociology classes to analyze patterns of North African demographics in France. These results were applied to the discussion in another French class reading short selections from francophone North African literature and about the political situation in France. The presentation will more fully describe the survey results, partially illustrated in Figure 4, based on the students' responses to questions concerning the team activity and their learning.

Applications of these maps to the study of French language and culture and francophone literature come into focus when one "zooms" into a region, city, Parisian *arrondissement* (district) and neighborhood and how it actually looks, much as Stendhal zoomed in on the little town of Verrières in the first chapter of Le rouge et le noir (1830). Our project constantly strove to put a human face on what might otherwise have simply been an aerial photo or a map showing colorful patterns. Zooming in on the La Glacière métro stop in Paris through French census data yields one sense of the community. Another one emerges when one views the neighborhood from street level and watches a documentary clip on street basketball (Figure 1).[4]

An historical example includes a GIS map of the population density in the Paris region which surprised us as an illustration compatible with the Concentric Zone Theory proposed by the Chicago School of Ernest Burgess, Robert Parks and others in the 1920's.[5] The results of the French policy of centralization start to become clear from this perspective by using a map that shows contrasts in population (Figure 3, originally in color). This map and similar ones have sparked student discussions of the differences in schooling between the United States and France, of the relative sizes of cities, and other conversations that have yielded a good number of "aha" moments. The complexity of these thoughts and their expression depend upon the students' language level. However, students starting with the second year of university French study have been able to grasp and discuss the implications of *"L'Etat, c'est moi"* for the general French population under Louis XIV, a statement that presaged the migratory movement that would lead approximately one quarter of the entire French population to live within a forty-four-mile (seventy-one-kilometer) radius starting from the center of Paris (Figure 3).

GIS tools have allowed us to create maps that help answer sociodemographic and other questions that might never have been asked. These new products have made visible certain patterns that would otherwise have been hidden in census and other data (Figure 4, originally in color). This presentation shares the background, materials and procedures by which the project was created and has been sustained in several disciplines since 1999. To our knowledge, no creation of similar materials for French and Sociology had been created.[6]

Further work will involve the collection of authentic materials by U.S. students on study abroad in France and by faculty members. The new materials will include interviews with residents whose lives may already span various generations, photographs linked to GPS coordinates, historical documents, and current realia. These authentic materials will be integrated into language acquisition curricula as well as higher-level courses by faculty in our French program in the proof of concept phase. One prospective benefit is the encouragement of study abroad as students become much more familiar with the people, sounds, sights, arts, thoughts and sociological fabric of various towns and cities in France. Quantitative and qualitative aspects will provide students with a fuller ability to appreciate French culture and civilization as well as give them the chance to work with peers who do not have the linguistic background to access the materials first hand. The results of a small-scale study of such interaction (Figure 5, originally in color) suggest that helping speakers of a foreign language add a GIS analytical ability at the university level and in a career is an easier task than the obverse, training users of GIS sufficiently in the foreign language to allow them first-hand access to the foreign language materials.

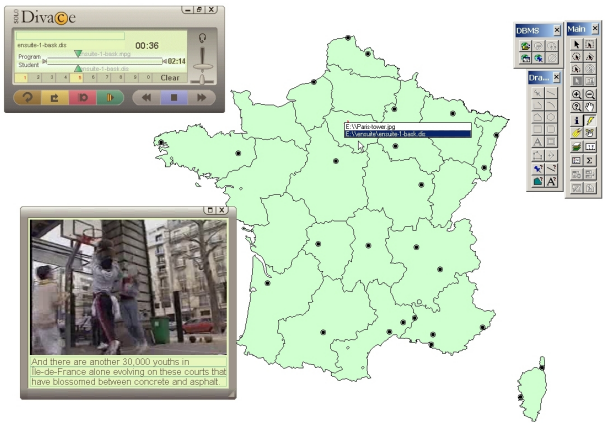


Figure 1:

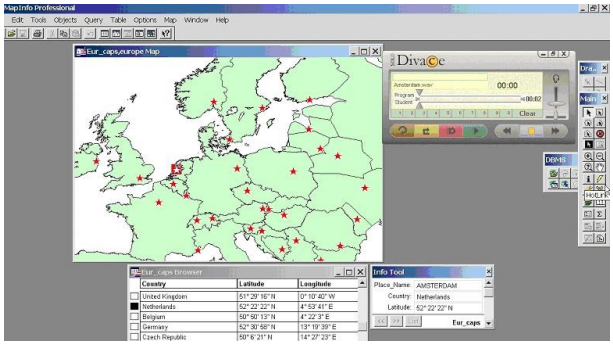


Figure 2:

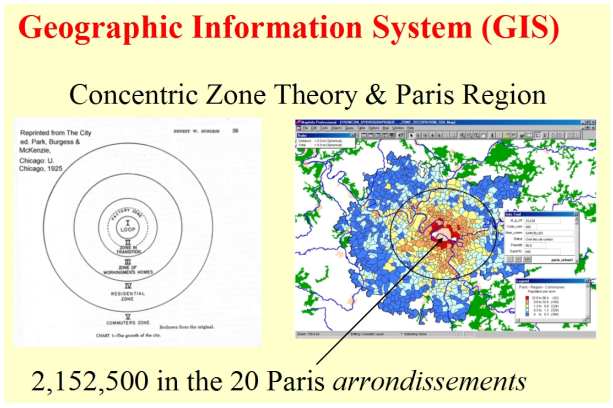


Figure 3:

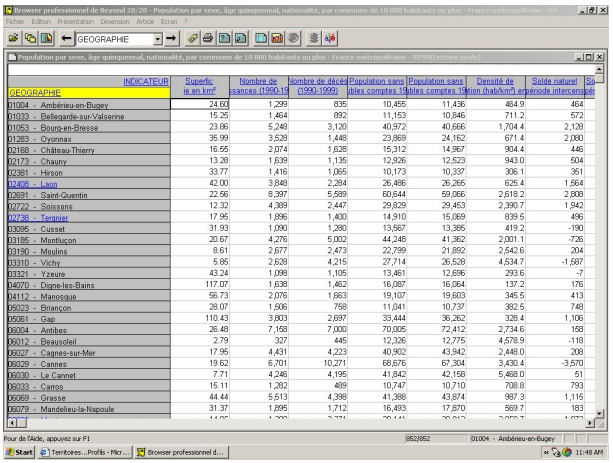


Figure 4:

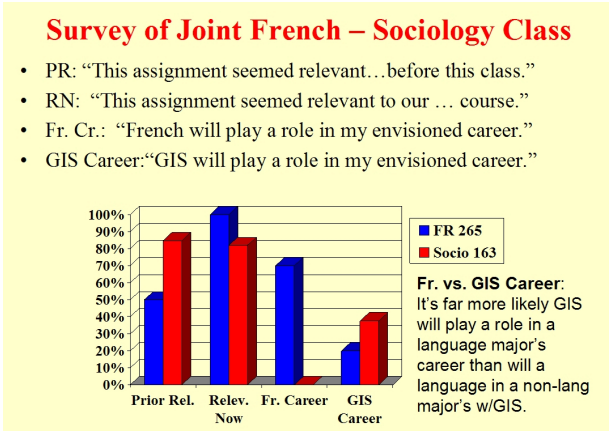


Figure 5:

---

1. See results reported on the Modern Languages, Literatures and Geographical Information Systems (GIS) Web page at <http://www.faculty.fairfield.edu/jgoldfield/MLL-GISprojects.htm> and <http://www.faculty.fairfield.edu/jgoldfield/ISLT-Webpg0201.htm> (Feb. 28, 2007). Dr. Kurt Schlichting, Chair and Professor Sociology and Anthropology at Fairfield University directed the ISLT Initiative and, together with Mr. Chris Calienes, assisted the author in the preparation of the GIS maps.
2. See information on successes of FLAC in the liberal arts and professional schools, such as at the University of Rhode Island and the University of Connecticut, at: <http://press.ucsc.edu/archive/95-96/10-95/101695-National_conference.html> (Feb. 28, 2007). An important report on FLAC from the American Council on the Teaching of Foreign Languages appears at: <http://www.actfl.org/files/public/Fall1995LangAcrs.pdf> (Feb. 28 2007)
3. The digital annotation tool was obtained from Tandberg Educational, Inc., and Divace Oy (now known as Sanako, Inc.),

Divace Solo, v. 4.0 (Turku, Finland: Divace Oy, 1997-2003). The product is now known as *Media Assistant Solo*, CD-ROM, published by Sanako, Inc. (Turku, Finland). The sample scene of inner-city basketball in Paris was part of instruction materials created by the author for French courses of various levels, necessitating varying types of linguistic assistance, such as the English subtitles pictured here. Other versions of the same audio-video materials included French captions or no linguistic support materials.

4. Thompson, Chantal P. and Bette G. Hirsch. Videotape to accompany *Ensuite: Cours intermédiaire de franc ais*. 4th ed. Boston: McGraw-Hill, 2003.

5. Burgess, Ernest. *The Growth of the City: An Introduction to a Research Project*. Eds. Robert Park, Ernest Burgess and R. D. McKenzie. Chicago: University of Chicago Press, 1925.

6. Since March 1, 2000, a monumental integration of GIS, historical, cultural, linguistic, artistic and other materials for the study of Asian cultures has been carried out by Professor David Germano and colleagues for the Tibetan and Himalayan Digital Library at the University of Virginia: <http://www.thdl.org/index.php> (Feb. 28, 2007).

Zeta and Iota and Twentieth-Century American Poetry

David L. Hoover *(david.hoover@nyu.edu)*
New York University

In his intriguing "All the Way Through: Testing for Authorship in Different Frequency Strata", John F. Burrows follows up his much-discussed Delta (Burrows, 2002a, 2002b, 2003; Hoover 2004a, 2004b, 2006) with two new measures of textual difference: Zeta and Iota (Burrows 2006; see also Burrows 2005).

Both measures begin with a full word frequency list for a sample of Restoration poetry (approximately 20,000 words) by a single primary author. The sample is then divided into five sections of equal size, and word frequency lists are created for them. Zeta deals with words of moderate frequency, words occurring in at least three of the five sections. To compare two poets, the word list is reduced further by removing any occurring more than three times in the second poet's sample. Where many authors are being compared, the list is reduced by removing any words present in the text samples of most of the other authors. Both methods remove from consideration the most frequent words of English that have been the focus of so much recent work. Whether there are two or many authors, the result is a list of words that are moderately frequent in the primary author but much less frequent in the other author(s).

For Iota, the word list is first limited to words appearing in at most two of the primary author's sections. To compare two authors, the list is further limited to words that are completely absent from the second author's sample. Where many authors are being compared, the list is further reduced by removing words that appear in more than half the other authors. In either case, very frequent and moderately frequent words are eliminated, leaving words that are not very frequent in the primary author but are rare or non-existent in the other author(s).

Zeta and Iota are remarkably effective in attributing poems as short as 1,000 words to the correct authors. Even more important, they allow the analyst to concentrate on a relatively small subset of characteristic words, nearly all content words. These lead back to the text and to important questions of interpretation and style.

Both Zeta and Iota will require further testing before they can be confidently applied to genuine questions of authorship and

style, and we can begin with a study of twentieth-century poetry. For these tests my corpus consists of samples of 14,000 to 129,000 words of poetry by twenty-six poets as the primary set and fifty-six independent poems from 900 to 21,000 words long as the secondary set, thirty-six of these by poets in the primary set and twenty by other poets (poems by primary authors are removed from their main samples). The texts were downloaded from Chadwyck-Healey's Literature Online and edited to regularize hyphens and to remove prose sections and non-authorial text, such as publication information, notes, section numbers, epigrams and other quotations. Delta tests were used to determine which of the poems and poets are most difficult to attribute, and these were analyzed using Zeta and Iota.

My head-to-head tests of Wallace Stevens vs Archibald MacLeish and Edwin Arlington Robinson vs Robert Frost give even more definitive results than Burrows achieves for Marvell vs Waller, though my much larger samples require minor adjustments in technique). The new measures have no difficulty distinguishing the two poets, whichever poet's word list is used.

When Burrows tests Marvell and Waller (using each poet's own primary word list) against the samples and twenty-four independent poems by twenty-four other main authors and twenty-one poems by other authors, Iota works very well for both authors, and Zeta works based on Waller's list. Marvell's list produces a group of failures which Burrows suggests are likely to be result from the contrast between the political satires being tested and the largely pastoral nature of most of Marvell's poetry.

My tests using the primary word lists of eight different authors yields strong results for Zeta on James Dickey, Vachel Lindsay, Robert Frost, and Wallace Stevens, with all of their individual poems ranking higher than any poem or sample by any other author. Zeta is very successful on two of William Vaughn Moody's independent poems, but the third ranks far below other author samples and individual poems, and it fails badly when based on the word lists of Edwin Arlington Robinson, Kenneth Rexroth, and Archibald MacLeish. For Iota, only the lists of Stevens and Lindsay produce completely correct results, though those of Moody and Rexroth produce good results except for a single poem by each author. Further research into the causes of these poorer results is underway, but the problems with Iota may be related to my larger samples. (The definition of "rare" is obviously very different for 20,000-word samples and 120,000-word samples.)

One alteration of Zeta that produces perfect results using Robinson's word list not only limits the word list to words that appear in at least three of the author's five sections, but also sets a lower limit on the word's frequency in the main set and limits the total frequency of the word in the twenty-five counter-sets. Another, still under investigation, calculates the standard deviation of the word's frequency in the five base sections and divides it by the mean frequency. Sorting the word list on the resulting Coefficient of Variation (or Relative Standard Deviation) makes it easy to limit the list to words that appear at relatively consistent frequencies in the five sections, besides appearing in at least three of them and in a limited number of the counter-set samples. A word appearing five times in each of a poet's five sections seems intuitively more "characteristic" of the author than one appearing twenty-three times in one section and once in each of two others.

Whatever the outcome of further testing and modification may be, Zeta, especially, is very effective in focusing attention on a poet's characteristic words, a useful task in its own right. In head-to-head tests on MacLeish and Stevens, much more stringent stipulations than Burrows used produce fascinating results: the twenty-six words occurring in all five sections of MacLeish's sample but with a frequency less than three in Stevens's sample are good potential MacLeish authorship markers, and the same stipulations produce forty potential Stevens authorship markers. These words range in rank within their word lists from about the 200th to the 1,400th most frequent. These two sets of marker words return our attention to the texts:

Characteristic Stevens words rare in MacLeish:
reality, except, centre, element, colors, solitude, possible, ideas, hymns, essential, imagined, nothingness, crown, inhuman, motions, regard, sovereign, chaos, genius, glittering, lesser, singular, alike, archaic, luminous, phrases, casual, voluble, universal, autumnal, café, inner, reads, vivid, clearest, deeply, minor, perfection, relation, immaculate

Characteristic MacLeish words rare in Stevens:
answered, knees, hope, ways, steep, pride, signs, lead, hurt, sea's, sons, vanish, wife, earth's, lifted, they're, swing, valleys, fog, inland, catch, dragging, ragged, rope, strung, bark

Stevens's words are longer and more abstract, especially the nouns, and his list is saturated with adjectives. MacLeish's list has very few adjectives and more verbs and concrete nouns. Searching for marker words in each poet's work yields a remarkable pair of short poems: MacLeish's short poem "'Dover Beach'–A Note to that Poem" (215 words) contains seven of his twenty-six marker words, including the three italicized in this brief passage:

> . . . It's a fine and a
> Wild smother to vanish in: pulling down---
> Tripping with outward ebb the urgent inward.
> Speaking alone for myself it's the steep hill and the
> Toppling lift of the young men I am toward now . . .

In his even shorter poem, "From the Packet of Anacharsis" (144 words), six of Stevens's forty marker words appear,

including the three italicized in this brief passage (internal ellipsis in the original):

> And Bloom would see what Puvis did, protest
> And speak of the floridest reality . . .
> In the punctual centre of all circles white
> Stands truly. The circles nearest to it share
> Its color . . .

Preventing the huge numbers of items being analyzed from masking any meaningful results is one of the most difficult challenges for quantitative analyses of literature. By selecting for examination words that are particularly characteristic of an author, Zeta and Iota are potentially very useful for literary analysis as well as authorship attribution, no matter what further refinements they may require.

Bibliography

Burrows, John F. "Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary & Linguistic Computing* 17.3 (2002a): 267-287.

Burrows, John F. "The Englishing of Juvenal: Computational Stylistics and Translated Texts." *Style* 36 (2002b): 677-99.

Burrows, John F. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37 (2003): 5-32.

Burrows, John F. "Who Wrote Shamela? Verifying the Authorship of a Parodic Text." *Literary & Linguistic Computing* 20.4 (2005): 437-450.

Burrows, John F. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary & Linguistic Computing* (2006). Advanced Access published January 6, 2006

Hoover, David L. "Testing Burrow's Delta." *Literary & Linguistic Computing* 19.4 (2004a): 453-475.

Hoover, David L. "Delta Prime?" *Literary & Linguistic Computing* 19.4 (2004b): 477-495.

Hoover, David L. "Word Frequency, Statistical Stylistics, and Authorship Attribution." *Advanced ICT Methods Guide to Linguistics.* Ed. Tony McEnery. : , Forthcoming.

Updating Delta and Delta Prime

David L. Hoover *(david.hoover@nyu.edu)*
New York University

John F. Burrows's Delta, a unitary measure of textual difference, has created a flurry of interest in authorship attribution and statistical stylistics since its introduction in 2001 (Burrows 2001, 2002a, 2002b, 2003, 2005; Hoover 2004b; García and Martín 2006; other studies are in progress). Burrows uses a Microsoft Excel spreadsheet to simplify and partially automate his calculation of Delta, and I have presented increasingly complex versions of The Delta Spreadsheet that more fully automate the calculation and the analysis of results (2004a, 2005a). I have also suggested possible improvements in how Delta is calculated or defined, which I have tentatively called "Delta Primes" (Hoover 2004c, 2005a, 2006).

The growing popularity of Delta analysis will surely lead to its further increased use. In addition to a number of studies in progress in the broad area of humanities computing, studies by other researchers are either already ongoing or being planned using Delta in the evaluation of high-stakes educational testing of writing in high schools, in the "authorship" of film direction, in evolutionary biology, and in studies of adverse drug reactions. The simplicity of automated versions of Delta analysis is critical to its increased use, especially for those who are not specialists in authorship attribution or humanities computing.

In its current state, The Delta Spreadsheet allows the user to paste in raw word frequency lists and then to perform a complete Delta analysis automatically by running a Visual Basic macro that calls other macros. One of the macros removes any words from the word list that are not found in the primary set of texts (the presence of such words would prevent the calculation of Delta by causing division by zero). Another macro changes the raw word frequencies into text percentages and inserts a zero frequency record in the word list for each text whenever any of the most frequent words does not occur in that text. This process is extremely tedious, time-consuming, and liable to error if done manually, especially on the large word lists that are now typically used in Delta analysis (800-4,000 words). Another optional macro removes personal pronouns. These pronouns are entered into a column in the spreadsheet, along with the master word list to be analyzed, and the user also can enter there any other words that should be eliminated (for example, noise words or proper names). The spreadsheet also allows the user to specify whether the word list should be culled to remove words for which a single text

provides most of the occurrences, and, if so, to specify what percentage of occurrences should be used as the cut-off. The user can also specify the size of the word list to create for the analysis and the total number of words to analyze; the analysis macro can also be set to run several times based on increasing or decreasing numbers of words. The result of all this automation is that an entire Delta analysis can be performed as a background task. And, on modest numbers of texts that are not very large, an entire analysis can be performed easily in an hour. This allows a researcher to try many different combinations of options in the search for the most accurate and reliable results.

My current project involves further elaboration of The Delta Spreadsheet to automate more of the necessary processes. I have created an additional spreadsheet into which the user can enter a list of texts to be processed. Once the texts and their authors have been entered and the primary and secondary text sets specified, the user runs a macro that collects the word frequency lists of these files from the current directory and adds the appropriate text and author labels. This allows the user then to cut and paste the word lists into The Delta Spreadsheet for processing, which simplifies the process, saves time, and reduces error. A third spreadsheet designed for quicker analysis of results allows the user to paste in the results of a series of Delta analyses and run a macro that reformats, sorts, and prepares the data for graphing, with similar benefits.

A final significant upgrade of The Delta Spreadsheet takes advantage of an analysis by Shlomo Argamon (forthcoming) of the statistical bases of Delta. Argamon shows that Delta (as well as my proposed Delta Primes) can be calculated without relying on the mean frequencies of words in the primary set of texts. The revised method of calculation makes possible a more streamlined version of The Delta Spreadsheet that allows me to increase the number of texts that the spreadsheet can process. It should also improve the performance of the macros, which can take several hours to process a large number of long novels. My poster will show how the modifications improve the performance of the spreadsheets and will present an example with a very large number of authors. I will also be prepared to do a live software demonstration of the operation of the spreadsheets at the conference, both on already existing word frequency lists and on texts which conference attendees supply.

Bibliography

Argamon, Shlomo. "Interpreting Burrows's Delta: Geometric and Probabilistic Considerations." *Literary & Linguistic Computing* (Forthcoming).

Burrows, John F. "Questions of Authorship: Attribution and Beyond." Paper presented at ALLC/ACH Joint International Conference, New York, June 14, 2001. 2001.

Burrows, John F. "Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary & Linguistic Computing* 17.3 (2002a): 267-287.

Burrows, John F. "The Englishing of Juvenal: Computational Stylistics and Translated Texts." *Style* 36 (2002b): 677-99.

Burrows, John F. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37 (2003): 5-32.

Burrows, John F. "Who Wrote Shamela? Verifying the Authorship of a Parodic Text." *Literary & Linguistic Computing* 20.4 (2005): 437-450.

Burrows, John F. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary & Linguistic Computing* (2006). Advanced Access published January 6, 2006

Garcia, A. M., and J.C. Martin. "Function Words in Authorship Attribution Studies." *Literary & Linguistic Computing* (2006). Advance Access published November 14, 2006

Hoover, David L. "Testing Burrows's Delta." Paper presented at ALLC/ACH Joint International Conference, Göteborg, Sweden, June 11-16, 2004.

Hoover, David L. "Testing Burrow's Delta." *Literary & Linguistic Computing* 19.4 (2004a): 453-475.

Hoover, David L. "Delta Prime?" *Literary & Linguistic Computing* 19.4 (2004b): 477-495.

Hoover, David L. "Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method." *ACH/ALL 2005 Conference Abstracts.* Victoria: University of Victoria Humanities Computing and Media Centre, 2005a. 83-84.

Hoover, David L. "The Delta Spreadsheet." *ACH/ALLC 2005 Conference Abstracts.* Victoria: University of Victoria Humanities Computing and Media Centre, 2005b. 85-86.

Hoover, David L. *The Delta Calculation Spreadsheet Online* . 2005c. <http://www.nyu.edu/gsas/dept/english/dlh/TheDeltaSpreadsheets.html>

Hoover, David L. "Word Frequency, Statistical Stylistics, and Authorship Attribution." *Advanced ICT Methods Guide to Linguistics.* Ed. T. McEnery. Forthcoming.

Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the *Encyclopédie*

Russell Horton *(russ@diderot.uchicago.edu)*
Digital Library Development Center
University of Chicago
Robert Morrissey
ARTFL Project
University of Chicago
Mark Olsen *(mark@barkov.uchicago.edu)*
ARTFL Project
University of Chicago
Glenn Roe *(glenn@diderot.uchicago.edu)*
ARTFL Project
University of Chicago
Robert Voyer *(rlvoyer@diderot.uchicago.edu)*
ARTFL Project
University of Chicago

One of the crowning achievements of the 18th century Enlightenment was the *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres*, edited by Diderot and d'Alembert. Published in Paris between 1751 and 1772, in 17 volumes of text and 11 volumes of plates, it contains 74,000 articles written by more than 140 contributors.[1]The *Encyclopédie* was a massive reference work for the arts and sciences, as well as a *machine de guerre* which served to propagate Enlightenment ideas. The impact of the *Encyclopédie* was enormous. Through its attempt to classify learning and to open all domains of human activity to its readers, the *Encyclopédie* gave expression to many of the most important intellectual and social developments of its time.

The scale and ambition of the *Encyclopédie* inspired its editors to adopt three distinct modes of organization which, taken together, Diderot described as encyclopedic: dictionary, hierarchical classification, and the *renvois* (cross-references). The interaction of these three modes has led modern commentators to describe the *Encyclopédie* as an "ancestor of hypertext" and to depict Diderot as "l'internaute d'hier"[2]. D'Alembert underscores the importance of the organization of knowledge in the *Discours Préliminaire*:

> As an *Encyclopedia*, it is to set forth the *order* and *connection* of the parts of human knowledge. As a *Reasoned Dictionary of the Sciences, Arts, and Trades*, it is to contain the general principles that form the basis of each science and each art ... and the most essential facts that make up the body and substance of each.[3]

Of the three modes of organization, the dictionary mode (organization of entries in alphabetical order) is certainly the simplest and the most arbitrary. The second mode of organization is classification, wherein each dictionary entry is assigned to a "class of knowledge," placing it within the "order" of human understanding, as depicted in the *Système Figuré des connaissances humaines*. Modeled after Bacon's classification of knowledge and Enlightenment theories of epistemology, all understanding is founded upon memory, reason, or imagination, with numerous categories and sub-categories branching out from these three faculties.[4] However, simply placing an entry into this hierarchy of knowledge was insufficient to indicate the interconnections of knowledge. Thus, Diderot created an extensive system of renvois, the third mode of organization, providing a lattice of interconnections between individual leaves of the tree as well as between classes of knowledge.[5]

The central role of the classification system in the intellectual objectives of the *Encyclopédie* editors is indicated by the extent to which it has been discussed and debated by both contemporaneous scholars and later researchers. The editors were remarkably diligent in assigning classes of knowledge to each article and sub-article. Of the 73,840 main and sub articles, 55,227 were assigned classes of knowledge. The editors were, however, somewhat less diligent in maintaining a precisely controlled list. Thus the classifications as found in the text are an amalgam of abbreviations, conflations, and even entries that are not found on the *Système Figuré*. We have recently completed orthographic normalization of the classes of knowledge assigned to each article,[6]" resulting in some 54,289 articles with 2,600 normalized classes of knowledge. The twenty most frequent classifications by number of articles are:

5513 Géographie
4794 Géographie moderne
3084 Géographie ancienne
2396 Jurisprudence
2304 Grammaire
1894 Marine
1483 Commerce
1277 Histoire naturelle. Botanique
1194 Histoire moderne
1115 Mythologie
1069 Histoire naturelle
889 Histoire ancienne
796 Medecine

730 Architecture
689 Jardinage
682 Littérature
627 Maréchallerie
614 Botanique
558 Histoire ecclésiastique
517 Théologie

Like the *Système Figuré*, these classifications are a reflection of how knowledge was ordered and classified in the 18th century. Given the assumptions that ontologies are historically contingent and that the *Encyclopédie* is by far the most consistent and coherent representation of the structure of 18th century knowledge in French, this paper reports the results of our current experiments using machine learning and data mining techniques to understand and exploit this unique resource. Our initial objectives are three-fold. First, we plan to examine the relationship of the classifications to the content of the articles using machine learning techniques to identify feature sets that characterize classes of knowledge in the 55,000 articles classified by the editors of the *Encyclopédie*. Secondly, we will apply these feature sets to the 19,500 articles for which we do not have a class of knowledge and evaluate the accuracy of classification by randomly selecting articles with known authors which scholars will then inspect. Most contributors to the *Encyclopédie* worked on fairly specific domains -- Rousseau contributed exclusively on music, for example -- so we can use authorship as one control for judging the accuracy of classification. Similarly, the cross-references will also serve as an evaluation control, since 50% of the *renvois* link to articles within the same class of knowledge. Finally, we plan to apply these feature sets to the unclassified "plate legends" in an effort to determine accuracy by examining the degree to which classification of the plate legends reflects the relationship of particular plate legends to particular articles.

For our initial experiments, we have extracted the text from all articles that are more than 100 words in length, and which are categorized within one of the 50 most frequent normalized classifications. Explicit markers of class of knowledge, present at the beginnings of these articles, were removed to ensure that they do not provide facile criteria for classification. The texts are tokenized and lemmatized[7], and frequencies of words and lemmas are computed both globally and for each article. Words and lemmas with more than 100 occurrences in the entire *Encyclopédie* were used as attributes, and vectors for each article were generated from the number of occurrences of each attribute in that article.

We are using the SMO implementation of a support vector machine in the Weka[8] data mining engine for initial experimentation on smaller data samples, and an SVM-Light[9] classifier for larger datasets. While support vector learning algorithms are very effective for classification problems[10], we are also evaluating several other data metrics and machine learning techniques, including information gain statistics and J48 decision tree classification as implemented in Weka, to examine the the most salient features that are used in the classification process and to test the effectiveness of various feature set selections.[11]

Results from preliminary experimentation indicate that SVM classifiers applied to the articles of the *Encyclopédie* are very effective at distinguishing articles from different classes. We examined 936 unlemmatized articles in our sample dataset belonging to the classes *Medecine* (499) and *Mythologie* (437). The Weka SMO classifier using default options with 10-fold cross-validation correctly recognized 98.29% of the articles (920/936). Under the same parameters, the Weka J48 tree classifier achieved slightly lower performance (91.66% accuracy). The decision tree showed a clear split on medical content words, such as *maladie*, *humeurs*, *inflammation*, and so on. Such strong performance may be due to the fact that one would not expect to find similar vocabularies in articles dealing with medicine and mythology. We achieved similar performance by classifying 2,448 articles equally divided between modern and ancient geography. The SMO training achieved 100% accuracy with 92.2% accuracy on cross-validation. Inspection of the most important features in both J48 tree and InfoGain measures shows a strong preference for classical authors (Pline, Ptolomée) and places (Gaule, Thrace), and the strings "l", "lib" and "liv", which correspond to citations of classical authors (e.g. Pline, l. IV. c. xvj.).[12] Distance and location terms (*lieues*, *long*., *latit*.) are strongly correlated with modern geography. Furthermore, the function words "*selon*" and "*dit*", which are far more prevalent in ancient geography articles as the authors were citing classical descriptions, are given high InfoGain scores.

We anticipate assigning classes of knowledge to articles that were not originally classified by the editors iteratively by comparing all unknown articles to specific classes of knowledge rather than trying to classify all unknown articles *en masse*. To test this approach, we assembled two sets of articles each containing 1,209 instances. The first set contained articles categorized by the editors as belonging to ancient geography, while the second set was constituted by selecting, for each article in the first set, an article as close as possible in length but belonging to a different class of knowledge. Using SMO training, we achieved 97.8% accuracy with standard 10-fold cross-validation. Again, the most heavily weighted features were terms denoting classical authors and place names, along with a greater preponderance of more general geographic terms, comporting nicely with a reasonable human's intuitive understanding of what makes a document on ancient geography distinct from another documents. We further validated our results by running another experiment identical to the first except that each article was randomly labeled as either ancient

geography or not ancient geography, irrespective of its true classification. The principle of Random Falsifiability states that if random labels can be learned with the same ease (for SVM, 'ease' can be defined as proportion of support vectors required [13]) as true class labels, the method must be rejected as unreliable. After 10-fold cross-validation, SMO achieved a mere 50.2895% accuracy on the classification, barely surpassing random chance. That our method cannot learn the random labels at all suggests that our success in discrimination is in fact based on inherent differences between the two classes and not merely a greedy model's exploitation of arbitrary patterns in the data distribution.

The SMO model derived from comparing ancient geography to a random selection of articles in other classes allows us to test classification on a set of unclassified articles. To do this, we assembled 5,000 randomly selected articles containing more than 100 words for which classification was unknown and with attributed authorship. We then applied the ancient geography SMO model to this set in an effort to identify articles pertaining to this category. The recall of this experiment was far too high. In the future we intend to implement a classifier that reports a numeric score rather than a simple binary categorization. There were, however, within the results a number of correctly classified articles such as the river *ASOPE* and the articles *GARAMANTES* and *Ionique Transmigration*. Many of the misclassified articles, such as *ADONIES*, *ou FESTES ADONIENNES* and *Danse astronomique*, pertain to classical history, mythology and other related fields. In addition to implementing a ranking classifier, we will also investigate moving up the tree of knowledge in order to use a coarser classification scheme; e.g., rather than remaining at the leaves of "ancient" and "modern" geography, we would use the branch of geography itself as a general category.

The impressive performance of machine learning algorithms suggests that the editors of the *Encyclopédie* were quite judicious in their assignments of classifications, a claim which will be tested further in the full paper. Examination of the features most effective in classification tasks will establish a sort of thesaurus which will give scholars a better understanding of the organization of knowledge during the Englightenment. Furthermore, we believe that the creation of well-verified training sets on this large corpus will allow us to test the degree to which we may profitably apply what the systems have learned to articles and plate legends which were not classified at the time, using the contemporary ontologies. If this series of experiments is successful, we would anticipate using the training sets from the classifications in the *Encyclopédie* to attempt to classify passages in other 18th century French documents.

---

1. The ARTFL implementation of the Encyclopédie is discussed in Robert Morrissey, Jack Iverson and Mark Olsen, "Présentation: L'Encyclopédie Electronique" Robert Morrissey and Philippe Roger, eds., *L'Encyclopédie de réseau au livre et du livre au réseau*, (Paris: Champion, 2001): 17-27, and Leonid Andreev, Jack Iverson and Mark Olsen, "Re-engineering a War Machine: ARTFL's Encyclopédie" *Literary & Linguistic Computing* 14.1 (1999): 11-28.
2. Eric Brian, "L'ancêtre de l'hypertexte", *Les Cahiers de Science et Vie* 47 (Oct. 1998): 28-38.
3. English translation cited in Nelly Hoyt and Thomas Cassier's "Introduction" to *Encyclopedia* (1965): xxiii (our emphasis).
4. For various representations of the Système Figuré and the Editors' description, see <http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/texts/> and <http://artfl.uchicago.edu/cactus/>.
5. Blanchard and Olsen examined the structure of the *renvois* generating a "mappemonde" of the cross-references and node level classes of knowledge. See Gilles Blanchard and Mark Olsen, "Le système de renvois dans l'*Encyclopédie*: une cartographie de la structure des connaissances au XVIIIème siècle", *Recherches sur Diderot et sur l'Encyclopédie* 31-32 (April 2002): 45-70.
6. This project was accomplished in collaboration with Professor Dena Goodman at the University of Michigan.
7. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>
8. Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques* 2nd ed. (Morgan Kaufmann, 2005) and <http://www.cs.waikato.ac.nz/ml/weka/>
9. SVM-Light: <http://svmlight.joachims.org/> See T. Joachims, "Making large-Scale SVM Learning Practical", *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola eds. (MIT-Press, 1999). Note that we are using a parallel implementation. See <http://www.dm.unife.it/gpdt/>, G. Zanghirati, L. Zanni, "A Parallel Solver for Large Quadratic Programs in Training Support Vector Machines", *Parallel Computing* 29 (2003): 535-551 and L. Zanni, T. Serafini, G. Zanghirati, "Parallel Software for Training Large Scale Support Vector Machines on Multiprocessor Systems", *JMLR* 7 (July 2006): 1467-1492.
10. S. Dumais, et. al., "Inductive learning algorithms and representations for text categorization", *CIKM-98*, 1998.
11. See the discussion conformity and uniformity in Chih-Ming Chen, et. al. "A Hierarchical Neural Network Document Classifier with Linguistic Feature Selection" *Applied Intelligence* 3 (December 2005).
12. We checked these using the PhiloLogic build of the *Encyclopédie* (<http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/>), suggesting the importance of checking text mining results with full text analysis systems.
13. A. Ruiz and P.E. López-de-Teruel, "Random Falsifiability and Support Vector Machines" (<http://learn98.tsc.uc

3m.es/~learn98/papers/abstracts/paper013/abstract.html>).

Understanding the Linguistic Construction of Gender in Shakespeare via Text Mining

Sobhan Raj Hota *(hotasob@iit.edu)*
Linguistic Cognition Lab
Illinois Institute of Technology
Shlomo Argamon *(argamon@iit.edu)*
Linguistic Cognition Lab
Illinois Institute of Technology
Rebecca Chung *(chung@iit.edu)*
Lewis Department of Humanities
Illinois Institute of Technology

1. Introduction

Can computational analysis better reveal how Shakespeare's words and phrases construct characters clearly gendered as male and female? What happens when stylistic analysis is brought to bear on a longstanding notion in literary and cultural studies that gender identity is a discursive (i.e. a culturally-decodable lexical and semantic) construction? How helpful can linguistic style be in explaining aspects of literary style?

Last year we presented our first results (Hota et al. 2006) in analyzing lexical items of character gender in Shakespeare. Our observations on gender character were in line with previous work (Argamon et al. 2003) on discriminating author gender in modern texts, supporting the idea that Shakespeare projects character gender in a manner consistent with patterns of authorial gender projection found in other texts both literary and nonliterary.

In this abstract, we extend and refine our methods by focusing on lexical and semantic language use by Shakespeare for determining the gender of his literary characters. Here we used a better version of Shakespeare corpus, the Nameless Shakespeare. In the Nameless Shakespeare, each lexical item is fully tagged with lemma entries. Our accuracy has improved to 81% using lemma features, compared to last year's results. We observe that lemmas and tri-grams help identify a Shakespearean character's gender. In addition, discrimination models using lemmas and tri-grams may allow literary and cultural scholars to see discursive patterns that, while impossible

to find noncomputationally, nonetheless portray or construct a character's attitudes as male or female

The fact that these patterns hold across literary and nonliterary texts, and from early-modern to modern English, supports their possible significance in understanding discursively-formed gender identity. We further observe other distinguishing features, including the fact that some feature constellations match well to previous reports of features that distinguish male from female authors (Argamon et al. 2003). We have analyzed the concordance lines of lexical and lemma tri gram occurrences from the corpus and found patterns of phrasal usage that indicate significant gender differences in language use in the plays.

We are interested in understanding gender characterization based on major characters and minor characters, and in understanding how prose and verse forms impact gendered speech in Shakespeare. We are also interested in how words highly-associated with a particular gender may also help with plotting, dramatic tension, and closure. We will present our findings at the conference.

2. Methodology

We applied text classification methods using machine learning with feature sets described above, under the umbrella of a well-tagged corpus. If reasonable classification accuracy is achieved with these new sets of features, it will show that Shakespeare used words differently for his male and for his female characters. If this is the case, then examining the most discriminating features should give some insight into how gender stylistics flows into socio-linguistic and literary gender construction.

2.1 Corpus Construction

We constructed a corpus of characters' speeches from 35 Shakespearean plays, collected from the lexically-and-lemmatically-tagged Nameless Shakespeare. The plays are in XML (Extensible Markup Language) format. To import them into our system, we extracted the speeches and the gender of each character automatically, cleaning the stage directions. A text file for each character in each play was constructed by concatenating all of that character's speeches in the play. We only considered characters with 200 or more words. From that collection, all female characters were chosen. Then we took the same number of male characters as female characters from a play, restricted to those not longer than the longest female character from that particular play. In this way, we balanced the corpus for gender, giving a total of 101 female characters and 101 male characters, with equal numbers of males and females from each play (see Table 1). We balance the corpus to avoid bias in the automated learning procedure; this introduces other issues which we address below.

Tables:

Play Name	Gender Count
All's Well That Ends Well	8
Antony and Cleopatra	6
As You Like It	6
Cymbeline	4
King Lear	6
Loves Labours Lost	8
Measure for Measure	6
Midsummer Nights Dream	8
Much Ado About Nothing	8
Othello The Moore of Venice	6
Pericles Prince of Tyre	6
Romeo and Juliet	6
The Comedy of Errors	8
The First part of King Henry The Fourth	4
The First part of King Henry The Sixth	6
The Life and Death of Julies Caesar	4
The Life and Death of Richard The Second	6
The Life and Death of Richard The Third	8
The Life of King Henry The Eighth	6
The Life of King Henry The Fifth	4
The Merchant of Venice	6
The Merry Wives of Windsor	6
The Second part of King Henry The Sixth	4
The Taming of the Shrew	4
The Tempest	2
The Third part of King Henry The Sixth	4
The Tragedy of Hamlet	6
Titus Andronicus	4
Troilus and Cressida	4
Twelfth Night	6
Two Gentlemen of Verona	6
Winter's Tale	6
The Tragedy of Coriolanus	6
King John	6
Macbeth	8

Table 1: Shakespeare Corpus

2.2 Feature Extraction

We processed the text using the ATMan system, a text processing system in Java that we have developed. The text is tokenized and the system produces a sequence of tokens. Each token corresponds to a word in the input text file. We used lexical and lemma features with n = 1, 2 and 3 gram combinations. In order to understand gendered language more deeply, we extracted those n-grams most linked to gender (Tables 2, 3). We collected most frequent 500 words, bigrams (2670) and trigrams (356) from the lexical entries. In the same way 2001 unigrams, 2860 bigrams, and 571 trigrams from lemmas were collected. We calculated the frequencies of these various features and computed their relative frequencies. The list of various feature sets with their counts is given Tables 3-5.

Features	Count
Function Words	646
500 Most Frequent Words	500
Bag of Words	2426
Bi Gram	2620
Tri Gram	356
Uni plus Bi Gram	5096
Bi plus Tri Gram	2976
Uni plus Tri Gram	2782
Uni plus Bi plus Tri Gram	5462

Table 2: Lexical Features

Features	Count
500 Most Frequent Words	500
Uni Gram	2001
Bi Gram	2880
Tri Gram	571
Uni plus Bi Gram	4881
Bi plus Tri Gram	3431
Uni plus Tri Gram	2572
Uni plus Bi plus Tri Gram	5432

Table 3: Lemma Features

2.3 Text Classification

The classification learning phase of this task is carried out by Weka's (Frank & Witten 1999) implementation of Sequential Minimal Optimization (Platt 1998) (SMO) using a linear kernel and default parameters. The output of SMO is a model linearly weighting the various features. Testing was done via 10 fold cross validation. With this methodology, we ensure that each character is tested on at least once with training that does not include it. Table 4 presents the results obtained by running various experiments.

Features	Lemma	Lexical
Function Words	-	62.37
Bag of Words	-	71.78
500 Most Frequent Words	80.69	76.73
Uni Gram	71.78	-
Bi Gram	63.36	67.82
Tri Gram	61.38	57.42
Uni Bi Gram	70.29	75.24
Bi Tri Gram	67.32	69.30
Uni Tri Gram	69.80	71.78
Uni Bi Tri Gram	72.27	75.74

Table 4: SMO Accuracy on Gender

3 Results: Accuracy and Feature Analysis

Many feature combinations give classification accuracies near or above 70%, which is quite good (random would be 50%, since the corpus is balanced). The highest accuracy of all (80.69%) was attained using the 500 most frequent word lemmas as features. Lexical (surface tokens) also worked well, with unigrams plus bi grams plus tri grams combination as a whole giving the highest accuracy (75.74%). The accuracy is captured in Table 4.

The feature analysis phase is carried out by taking the results obtained from Weka's implementation of SMO. SMO provides weights to the features corresponding to both class labels. To discriminate binary class labels, SMO uses positive and negative weight values in a linear model. After sorting the features based on their weights, we collected the top ten features indicative of each gender. We have also computed the average value of each feature for each gender (Tables 5-10). For reasons of space we consider here just those feature sets giving the most insight (as well as good classification accuracy). We also ranked features using information gain (IG), defined as the expected reduction in entropy caused by partitioning the training set according to the attribute.

Lemma Unigrams:
In Shakespeare, several meaningful clusters of words emerge. Female lemmas indicate family relationships ('husband', 'mother', 'court') , feelings ('sick', 'merry'), emotional injections ('alas', 'o', 'prithee'), and integration of personal context ('he', 'you'). Male features indicate concern with quantification ('three') and social status ('noble', 'solemn', 'savage'). Male lemmas also include some less-clearly interpretable verb forms ('begin', 'alight', 'beat').

Lemma Trigrams:

More specific meaning patterns can be seen in lemma triples. Female trigrams mostly indicate construal of self and others (‘I/see/you’, ‘for/I/to’, ‘I/know/I’, ‘be/he/not’, ‘say/I/be’), politeness (‘thank/you/for’), conditionals (‘if/he/have’), and questions (‘who/be/that’). Male trigrams focus on assertions (‘I/say/to’) mainly about personal/social status (‘but/I/be’, ‘be/a/very’, ‘be/a/ass’, ‘I/be/ he’), possessions (‘have/no/more’, ‘I/have/lose’), and manner (‘the/manner/of’).

Analysis of Concordance Lines:

For both lexical and lemmatic trigrams, we contextualize usage by examining concordance lines. For males, 'the name of'’ is followed usually by 'truth', 'hero', 'love', 'justice', 'whore', while for females, this trigram is followed by 'wife', 'jesting' and sometimes with the name of a female character. Men use this to invoke overarching abstractions (or to insult women), while women talk about “names of” in a less metaphorical and more neutral fashion. For 'do you know', males tend to follow with another question, but not females. Strikingly, males use 'the manner of,' to set up dramatic contrasts or even shifts in dramatic action, but female use of this trigram is entirely unremarkable in this way.

Top-10 Gender Character Features - SMO Weight and Average Frequency(x100)

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
being	0.9219	70.79	44.05	never	-0.9578	112.87	90.59
well	0.8956	245.54	179.70	such	-0.8877	144.55	122.77
already	0.836	19.30	7.92	little	-0.7257	53.46	41.08
we	0.7607	828.21	649.00	yet	-0.6886	157.42	139.60
thank	0.7242	53.96	43.56	only	-0.6845	30.19	27.72
allow	0.7184	7.92	3.96	wish	-0.6632	37.62	19.80
many	0.7123	55.94	40.09	hence	-0.6143	37.62	25.74
doing	0.7105	9.90	4.45	take	-0.6096	126.23	132.67
whom	0.6451	49.50	19.80	comes	-0.5989	50.49	49.00
there	0.6385	267.82	203.46	you	-0.5959	2054.45	2050.99

Table 5: FWs - All

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
whom	0.2084	49.50	19.80	Alas	-0.2558	50.0	13.36
three	0.1821	46.53	15.34	O	-0.1957	3334.15	2658.91
her	0.1651	655.44	465.34	Gone	-0.1832	57.92	36.13
fellow	0.1579	46.03	20.79	Grieve	-0.1614	10.89	3.96
degree	0.1572	5.94	1.98	husband	-0.1598	98.01	19.80
lying	0.1482	4.95	0.49	Heart	-0.1526	122.27	96.03
knit	0.1481	6.43	1.98	He	-0.1526	1689.60	1560.89
beat	0.1414	27.22	7.42	Prithee	-0.1472	35.64	13.36
avoid	0.141	5.44	2.47	Knife	-0.1453	5.94	2.47
to	0.1394	2167.82	1794.05	Sick	-0.1364	26.23	11.38

Table 6: BoWs – All

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
begin	0.2297	28.21	12.37	alas	-0.2864	50.0	13.36
three	0.2293	46.53	15.34	O	-0.2277	3321.78	2701.98
alight	0.1988	5.94	0.49	husband	-0.2236	98.01	19.80
solemn	0.1904	8.41	4.45	prithee	-0.2213	35.64	13.36
to	0.1867	2159.90	1778.71	mother	-0.1718	56.43	24.25
noble	0.1831	69.80	38.61	court	-0.1711	40.09	37.62
who	0.1754	248.51	165.84	he	-0.1707	2349.01	2326.73
she	0.173	774.75	559.90	you	-0.1667	2156.43	2145.54
beat	0.1721	28.21	7.42	sick	-0.1636	26.73	11.38
savage	0.1685	5.94	2.47	merry	-0.1634	23.76	10.89

Table 7: Lemma (Uni gram) - All

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
who	0.7804	248.51	165.84	alas	-0.8477	50.0	13.36
begin	0.6433	28.21	12.37	court	-0.6075	40.09	37.62
she	0.6331	774.75	559.90	you	-0.6074	2156.43	2145.54
three	0.6157	46.53	15.34	prithee	-0.5705	35.64	13.36
noble	0.5745	69.80	38.61	husband	-0.5585	98.01	19.80
this	0.4985	659.40	469.80	merry	-0.5466	23.76	10.89
well	0.4878	245.04	179.70	o	-0.5345	3321.78	2701.98
fellow	0.4831	45.54	22.27	mother	-0.5319	56.43	24.25
beat	0.4728	28.21	7.42	wish	-0.481	38.61	19.80
knave	0.4661	31.18	9.90	false	-0.4614	43.56	26.23

Table 8: 500 Most Frequent Lemma - All

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
but i be	0.9538	8.41	4.45	if he have	-0.8425	4.45	1.98
be a very	0.7805	4.95	1.48	thank you for	-0.7718	4.95	0.99
be a ass	0.7053	5.44	0	i see you	-0.6734	4.95	2.97
be to be	0.6776	6.93	0.99	for i to	-0.6616	4.95	2.47
have no more	0.6746	4.95	0.99	one of you	-0.6265	5.94	3.96
i be he	0.6514	9.40	2.97	i know i	-0.6081	7.42	5.44
i say to	0.6327	4.45	0.49	who be that	-0.5925	4.95	0
i have lose	0.628	4.45	0.99	be he not	-0.5602	5.94	0.99
the manner of	0.6249	5.94	1.98	say i be	-0.5354	4.45	0.99
i go to	0.6191	3.96	1.48	i be you	-0.5167	11.88	4.95

Table 9: Lemma (Tri grams) – All

Male Features				Female Features			
Feature	SMO Weight	Avg. Freq. (x 100)		Feature	SMO Weight	Avg. Freq. (x 100)	
		Male	Female			Female	Male
as much as	1.354	6.43	0.99	i warrant you	-1.1851	7.42	2.97
is to be	1.1555	6.43	0.49	what is it	-1.1449	4.95	1.48
the name of	1.1366	7.92	2.97	he is not	-1.0918	8.41	6.93
is a good	1.0875	5.94	1.48	i should be	-1.0458	4.95	1.98
away with him	1.0436	3.46	1.48	i am a	-0.9823	30.19	26.73
do you know	0.9901	4.95	2.97	you for your	-0.9625	4.45	2.97
not in the	0.9623	5.44	1.98	is it not	-0.8855	3.96	1.98
you are my	0.9462	3.96	1.48	by your leave	-0.8785	2.97	3.96
is but a	0.9395	6.93	2.47	for me to	-0.8754	4.45	2.47
it was not	0.9392	3.96	0.99	i am your	-0.8424	7.92	2.47

Table 10: Lexical (Tri grams) – All

Gender Character Features - Information Gain (IG) Weight and Average Frequency(x100)

Feature	IG Weight	(Avg. Frq. x) 100 Male	(Avg. Frq. x) 100 Female
is to be	0.0461	4.95	0
that i may	0.0415	3.96	9.90
will you go	0.0408	0.99	6.93
you are my	0.0356	3.96	1.48
to the king	0.0356	8.41	2.47
for mine own	0.0356	4.45	0.49
is not this	0.0304	3.96	0.99
me thy hand	0.0304	5.44	0.49
and we will	0.0304	3.96	1.48
i'll to the	0.0304	3.96	0
but i am	0.0304	3.46	1.48

Table 11: Lexical (Tri grams on IG) – All

Feature	IG Weight	(Avg. Frq. x) 100 Male	(Avg. Frq. x) 100 Female
i know you	0.0567	6.43	11.38
i to the	0.0463	14.35	4.45
to the king	0.0461	8.91	2.47
who be that	0.0461	0	4.95
if there be	0.0408	5.94	2.47
and we will	0.0408	4.95	1.48
will you go	0.0408	0.99	6.93
say i be	0.0408	0.49	4.45
be to be	0.0367	6.43	0.99
i love you	0.0356	0.99	5.94
i say to	0.0356	4.45	0.49
for i own	0.0356	4.45	0.49
be a ass	0.0356	5.44	0
i have lose	0.0304	4.45	0.99
and give i	0.0304	0.99	3.96
be a good	0.0304	6.43	0.99
where be i	0.0304	1.48	5.44
i thou hand	0.0304	5.44	0.49
be there no	0.0304	3.46	1.48

Table 12: Lemma (Tri grams on IG) – All

4 Discussion

These findings capture word patterning in Shakespeare inaccessible to non-computational methods of literary analysis, because of the scale of data processing involved. Literary scholars work almost exclusively with well-elaborated methods of semantic analysis (New Criticism, structuralism, and post-structuralism), developed with all the strengths and limitations posed by a book-only, eye-centered, subjectivity-dependent research context. In contrast, these findings encourage comparisons between non-computational and computational approaches. It is remarkable that these findings support aspects of non-computational methodology (words linked together in meaningful patterns like informational discourse/male and involved discourse/female), while also bringing to light new structural features of Shakespeare's discursive gender construction through language: parts of speech use, tri-gram combinations of words. These findings may, in addition, capture creative and literary patterning in greater detail than is possible with noncomputational literary methods alone. Since Shakespeare's plays depend greatly on gender-identified characters, the words linked to gender most likely also serve literary purposes. In the plays, heterosexual romance is linked not only to characterization, but also to action, and in the case of romantic comedies, to how the plays conclude. In the cases of complex characterization, a character's misuse of gendered speech may also be central to how Shakepeare develops dramatic action.

Limitations:

The editorial procedures for The Nameless Shakespeare are sound and practical for the project's purposes and for the work of this paper, but they need to be read and understood fully: both by literary scholars wanting to apply these findings to particular words in particular plays, and by computational scholars thinking through the problem of establishing textual accuracy prior to inviting a wider community to conduct searches. Also, with respect to Shakespeare's literary art, the findings here do not at this stage account for the impact of blank verse dialogue (for high or elite characters) versus prose dialogue (for low or common-born characters) on word choices and the numbers of words. It may be that blank verse fosters semantically significant tri-gram constructions because Shakespeare needed short words to complete plays primarily written in ten-syllable lines. But at least the question can be asked, and the answer will tell us something about Shakespeare both as dramatist and as poet. In addition, our work focuses on the heternormativity clearly present in Shakespeare's plays, but does not exclude nonheteronormative gender construction. Finally and significantly, the gender-balancing used here had the odd result of excluding from the corpus all major male characters in Shakespeare, including every male character named in a play's title, because all these characters speak more than 600 lines. All major female characters (including females named in titles) are included. We now know that very-long speech length per play efficiently identify characters as male, but we also will test our findings on the males excluded so far and report the results at the conference.

5 Conclusions

This is the first work, to our knowledge, in analyzing various textual features (lexical and lemma) collected from a single source in understanding literary character gender. We see, as in our earlier work (Hota et al. 2006) that the male and female language in Shakespeare's characters is similar to that found in modern texts by male and female authors (Argamon et.al 2003). Here we also observed the importance of trigrams for lexical and lemma features. Trigrams are few in number, so they are information rich and computationally efficient for identifying gender. The true import of the features identified by this analysis need to be confirmed by more traditional digital humanities methods such as examining concordance lines, to allow a more properly contextual interpretation. In any case, we believe that this study shows how classification learning can be used as a tool in developing new 'statistical' interpretative methodologies for bodies of literary works.

Acknowledgements:

Many thanks to Dr. Martin Mueller for providing us the Nameless Shakespeare corpus and many helpful comments in gender characterization in Shakespeare.

Bibliography

Argamon, Shlomo, Moshe Koppel, and Galit Averni. "Routing Documents According to Style." *Proceedings of the First International Workshop on Innovative Internet Information Systems (IIIS-98)* . 1998.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. "Gender, Genre, and Writing Style in Formal Written Texts." *Text* 23.3 (2003): 321–346.

Corney, Malcolm, Olivier de Vel, Alison Anderson, and George Mohay. "Gender Preferential Text Mining of E-mail Discourse." *Proceedings of 18th Annual Computer Security Applications Conference ACSAC* . 2002.

Hota, Sobhan, Shlomo Argamon, Moshe Koppel, and Iris Zigdon. "Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters." *Digital Humanities 2006 Conference Abstracts*. Paris: CATI, Université Paris-Sorbonne, 2006. 100--106.

Joachims, Thorsten. " Text Categorization with Support Vector Machines: Learning with Many Relevant Features." *ECML-98, Tenth European Conference on Machine Learning*. 1998.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. "Automatically Categorizing Written Texts by Author Gender." *Literarcy & Linguistic Computing* 17.4 (2002): 401-12.

Mueller, Martin. "The Nameless Shakespeare." *TEXT Technology* 14.1 (2005): 61-70. <http://texttechnology.mcmaster.ca/pdf/vol14_1_06.pdf>

Platt, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Microsoft Research Technical Report MSR-TR-98-14. 1998.

Witten, Ian, and Eibe Frank. Weka3: Data Mining Software in Java. <http://www.cs.waikato.ac.nz/ml/weka/>

Distinguishing Editorial and Customer Critiques of Cultural Objects Using Text Mining

Xiao Hu *(xiaohu@uiuc.edu)*
University of Illinois at Urbana-Champaign
J. Stephen Downie *(jdownie@uiuc.edu)*
University of Illinois at Urbana-Champaign
Andreas Ehmann *(aehmann@uiuc.edu)*
University of Illinois at Urbana-Champaign

1. Introduction:

There exists a large number of critical writings regarding humanities objects such as reviews, forum posts, mailing lists and blogs. In many cases, readers do not necessarily know the authenticity and credibility of such writings. It is desirable to have a tool that is able to distinguish professional criticisms from lay comments, and furthermore, to measure the authenticity of criticisms on humanities objects. Such tools can have many applications ranging from mass mail filtering to customized criticism recommendation and summarization. In a preceding study we demonstrated that a simple machine learning model can be used to automatically differentiate *editorial* critiques (i.e., those written by professional critics) from *customer* critiques (i.e., those written by interested members of the general public) (Hu et al. 2006a). In this poster, we extend and build upon our earlier work to include a new class of cultural objects (i.e., United States Literature) and to uncover the set of influential features that contribute to making "editorial" and "customer" reviews distinct.

For the sake of comparison, we use the same dataset as in (Hu et al 2006a), namely reviews from amazon.com, the largest online retailer of various humanities materials including books and music. On amazon.com, many book and music objects have both editorial reviews and customer reviews. The former are written by editors in amazon.com, who can be seen as experts, while the latter are written by arbitrary users from the general public. Besides the two product categories analyzed in (Hu et al 2006a), British Classic Literature books and Classical music CDs, we add a third category, United States Classic Literature. The three categories are among the most relevant to the humanities, yet cover different media and cultures. To eliminate possible product bias, we downloaded both editorial and customer reviews of the same objects that were randomly selected through the Amazon Web Services open APIs[1]. The descriptive statistics are shown in Table 1. It is noteworthy that the length of the customer reviews is highly variable. Also, customer reviews on classical music have a smaller vocabulary than editorial reviews.

Table 1: Statistics of experiment datasets

	British Literature		Classical Music		U.S. Literature	
Review Categories	Editorial	Custom	Editorial	Customer	Editorial	Customer
Number of Reviews	1425	1425	1843	1843	2203	2203
Total number of tokens	165,028	332,53	268,244	317,312	315,854	433,691
Term list size (tokens)	18,303	24,991	26,518	24,693	29,742	32,189
Total number of function words	76,659	180,31	126,543	170,283	152,711	243,503
Function word list size	388	415	394	414	403	415
Mean of review length (tokens)	115.81	233.36	145.55	172.17	143.37	196.86
Std Dev of review length (tokens)	83.13	215.31	77.76	180.68	95.76	185.09

Table 1

2. Experimental Setup:

Hu (Hu et al 2006a) demonstrated that a binary text classifier based on a Nai ve Bayesian model was able to classify editorial and customer reviews at accuracies of about 86%. Two sets of features were used in building the models. The first one is unigrams of all original tokens including content words, function words and punctuations, with the purpose of preserving all stylistic clues carried by the original writings. The second one is unigrams of all function words[2] which can carry important stylistic fingerprints (Stamatatos et al., 2002, Argamon et al., 2003). We discovered interesting features unique to each type of criticism. For example, professional critiques tend to use numbers while customers tend to use terms referring to personal experiences. However, unigram features bear virtually no context information. Therefore in this poster, we deepen our understanding by analyzing features with broader context information: bigrams and trigrams of the aforementioned two sets of tokens (Banerjee & Pedersen 2003). For easy comparison, the results of the classification experiments are presented in Table 2. It shows function words, though a small token set, can capture most of the differences of the two kinds of criticism. Comparing features with varying depths of context, we found bigram features consistently improve classification accuracies (to a level of 87.22% -- 89.25%) while trigram features do not achieve consistent improvement on classification accuracies. It is noteworthy that trigrams of function words are not as good as other feature sets. In fact, the genuine function word trigrams are too sparse in the datasets, so here we define a function word trigram is a sequence of 3 function words that occur within a window of 5 tokens in the text.

Table 2: NB classification on editorial and customer reviews*

Features	British Literature	U.S. Literature	Classical Music
Unigrams of all tokens	86.71% (1.90%)	86.86% (1.80%)	85.97% (1.65%)
Unigrams of function words	83.81% (3.21%)	83.73% (1.90%)	85.41% (1.60%)
Bigrams of all tokens	89.16% (0.84%)	89.16% (1.41%)	86.95% (1.65%)
Bigrams of function words	81.92% (2.01%)	81.19% (1.84%)	81.68% (2.54%)
Trigrams of all tokens	72.80% (1.50%)	83.55% (1.73%)	87.78% (1.36%)
Trigrams of function words**	66.99% (2.94%)	72.29% (1.91%)	76.32%(2.23%)

**Measures are mean accuracies of 10-fold cross-validation, in parentheses are standard deviations of accuracies.*
*** Trigrams of functions words are calculated within windows of size 5.*

Table 2

Upon a closer examination of the classification result on British Literature review set using trigrams of all tokens where the mean accuracy (72.80%) significantly worsened compared to its bigram and unigram counterparts, we found only 1.1% editorial reviews were wrongly classified as customer reviews but 53.1% customer reviews were misclassified as editorial reviews. This means the model can identify features unique to customer reviews but cannot reliably identify those unique to editorial reviews. A possible reason for this phenomenon is the British Literature dataset is dominated by reviews of similar books. This is supported by the bigram and trigram feature analyses described in next section. A large portion of the top features in this dataset is related to Shakespeare.

3. Feature Analyses:

Knowing that editorial and customer reviews are separable is not sufficient in and of itself; rather, we must strive toward understanding what features make them distinct. The binary Nai ve Bayesian text classifiers applied here can rank the terms according to their relative importance in the construction of the categorization model (Hu & Downie 2006). Table 3 – 8 list the top 10 features from each of the six feature sets in each review categories.

Table 3: Top unigrams of all tokens

British Literature		U.S. Literature		Classical Music	
Editorial	Customer	Editorial	Customer	Editorial	Customer
Brontë	quot	CliffsNotes	quot	Bartók	quot
CliffsComplete	definitely	eBook	recommen	Includes	_
exhibitions	recommend	Paperback	negative	adapted	bought
eBook	Another	Resource	depressing	bluegrass	reviewer
magnet	worth	Stavans	EH	songwriter	am
Scene-by-scene	boring	Includes	tedious	shrewdly	Takacs
Freshly	Helena	glossaries	mentioned	Tynan	cd
Edited	maybe	charged	basically	Takács	el
Bibliography	Overall	award-winning	NOT	scorer	me
Illustrations	actually	Boswell	re-read	bassist	purchased

Table 3

Table 4: Top unigrams of function words

British Literature		U.S. Literature		Classical Music	
Editorial	Customer	Editorial	Customer	Editorial	Customer
eighty	definitely	fiftieth	i'd	resulting	i'd
unto	maybe	fifteenth	i'm	unusually	i
million	i'd	twentieth	due	unlikely	am
eighteenth	actually	sixty	afterward	million	i've
hundredth	won't	forty	prior	tomorrow	myself
forty	i've	seventy	i	seldom	my
whence	i'm	nineteenth	moreover	occasionally	i'm
seventh	you're	hundredth	definitely	everywhere	therefore
thousand	that's	she'll	done	eight	besides
seventy	someone	sixteenth	i've	similarly	unless

Table 4

Table 5: Top bigrams of all tokens

British Literature		U.S. Literature		Classical Music	
Editorial	Customer	Editorial	Customer	Editorial	Customer
An essay	I found	study guides	I found	Academy Award	I love
Each edition	! !	CliffsNotes	. quot;	. Includes	! I
edited text	. quot;	Times Book	. quot;	, respectively	! !
Shakespeare Library	. quot;	new introduction	I read	Award winning	I bought
Folger Shakespeare	my favorite	eBook has	recommend this	II :	I really
an Introduction	recommend this	Drawing on	of quot;	. Conductor	recommend this
of rare	that there	recognized as	the quot;	adaptation of	have heard
globe .	read this	Critical Editions	I enjoyed	an orchestral	this cd
Library's vast	a bit	Each volume	I also	Ninth Symphony	time I
conveniently placed	I read	also features	you like	ambient music	my favorite

Table 5

Table 6: Top bigrams of function words

British Literature		U.S. Literature		Classical Music	
Editorial	Customer	Editorial	Customer	Editorial	Customer
of rare	that there	them again	i also	into their	as i
from around	if you're	sometimes be	you like	there's more	but i
to two	you are	as never	you should	if he's	so i
since its	so i	at more	as good	and	i am
you'll also	you get	how does	good as	is sure	which i
at more	he does	you'll also	while i	once upon	i also
into its	get the	and next	though i	for whom	i must
or another	is actually	than forty	i first	certain to	i have
can sometimes	that you	of six	the us	might as	i don't
and art	i can	twenty or	but i	again that	i like

Table 6

Table 7: Top trigrams of all tokens

British Literature		U.S. Literature		Classical Music	
Editorial	Customer	Editorial	Customer	Editorial	Customer
Folger Shakespeare Library's	. quot; The	CliffsNotes study guides	. quot; The	. Conductor Quintet	! ! !
Folger Shakespeare Library	. and we	of the western	. quot; is	, and expression	, but I
on the best	. I would	Times Book Review	, but I	of the Fates	. I am
the Folger Shakespeare	. The plot	, and John	I read	of the Mozart	. quot; and
the Folger offers	. I found	in the twentieth	is a very	, and Chopin	, I think
; An essay	in the book	York Times Book	is a good	of the stereo	, I have
; An introduction	. I don't	study guides offer	. I also	, and always	I have heard
Shakespeare Library in	! ! !	, The Bells	, I found	in the 1970s	. I think
for Shakespeare scholars	I have read	Penguin Classics is	. . quot;	original Broadway cast	. quot; is
around the globe	, and if	a new introduction	to read it	II : Larghetto	. I also

Table 7

Table 8: Top trigrams of function words (within windows of size 5)

British Literature		U.S. Literature		Classical Music	
Editorial	Customer	Editorial	Customer	Editorial	Customer
around the in	if you are	here the first	is a good	he with a	i have to
from around in	i have to	here for first	is a very	to his a	i was to
on the early	i it to	for first the	i would this	way with the	i am a
for from around	if you have	you'll it all	the i was	as a certain	i like the
of rare by	but it a	from both of	if you like	on such as	i was the
an by an	to get the	since its in	i this to	as a is	i to it
from around the	he does not	nearly after his	the is very	well as by	to the i
sometimes be to	the if you	nearly forty his	is not as	as his and	and i was
that can sometimes	this was a	nearly forty after	i had the	the to their	i that the
that sometimes be	and i have	in the twentieth	the and i	the but his	i am to

Table 8

4. Discussion:

As we can see from the tables, there are features consistent in editorial reviews across the three types of humanities objects that distinguish editorial critiques from customer critiques:

1. Numbers, both ordinal and cardinal, e.g., “in the twentieth”, “than forty”, “hundredth”
2. Technical terms e.g., “songwriter”, “bassist”, “in D Major”
3. Author or artist names using diacritical characters: e.g., “Bronte ”, “Barto k”, “Taka cs"
4. Authoritative resources, e.g., “Folger Shakespeare Library”, “CliffsNotes study guides”
5. Emphasis of the third person voice

Similarly, there are important features found in the customer reviews that contribute to their identity:

1. Terms referring to personal experience and the first person voice: e.g., “I found”, “I’m”, “I read”
2. Exclamation marks (“!”): from unigram to trigram, “!” consistently appear as top features.
3. Adverbs: e.g., “definitely”, “actually”, “possibly”
4. Contractions: e.g., “I’d”, “won’t”, “you’re”, “that’s”, “wouldn’t”,
5. Variations of artist names without diacritical characters: e.g., “Bartok”, “Takacs”
6. Quotations (“"” in XML documents) : e.g., “. quot; The ”, “, quot; is”
7. Colloquial phrases, e.g., “is like”, “is actually”, “have to say
8. Nonstandard words and marks: e.g., “_”, “cd”, “cds”

Most of the observed differences seem reasonable. Experts like to have more accurate descriptions by using numbers, citing authoritative resources, etc. While experts write in a more objective manner by using a third person voice, ordinary readers tend to connect humanities objects with their own personal experiences and prefer to express their emotions. Experts also use many technical terms while ordinary readers tend to use informal writing styles such as spoken language, contractions and nonstandard marks. Both experts and common readers refer to authors or artists, but very few readers bother using proper diacritical characters, instead opting to use basic Latin letters. It is interesting to see that common readers use more quotations, adverbs and punctuations than experts.

5. Conclusions and Future Work:

We extended our previous work on classifying editorial/professional reviews and customer reviews on humanities objects, and then examined the influential features in each of the review categories. Particularly, we examined two feature sets, “all tokens” and “function words only”, with context depths ranging from unigrams to trigrams. Results show that the two kinds of reviews are distinct. By using the NB feature ranking method, we found interesting and unique features associated with each type of review. Such features are important in enriching digital humanities repositories and facilitating criticism filtering and recommendation. The feature analyses also disclose new questions such as how criticism on literatures from various countries (e.g. British vs. U.S.) differs among one another. This will be part of our future work. We will also examine other critical writing resources such as mailing lists and forums for similar patterns between “expert” and “lay” contributors.

---

1. aws.amazon.com
2. The list of function words was edited by the Laboratory of Linguistic Cognition at Illinois Institute of Technology. It is available at <http://shekel.jct.ac.il/~argamon/gender-style/function-words.txt>

Bibliography

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. "Gender, Genre and Writing Style in Formal Written Texts." *Text* 23.3 (2003): 321–346.

Banerjee, Satanjeev, and Ted Pedersen. "The Design, Implementation, and Use of the Ngram Statistic Package." *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Feb. 2003.* Mexico City, Mexico, 2003.

Hu, Xiao, and J. Stephen Downie. "Stylistics in Customer Reviews of Cultural Objects,." *Proceedings of the 2nd SIGIR*

2006 Stylistics for Text Retrieval Workshop, Aug. 2006, Seattle, Washington. 2006.

Hu, Xiao, J. Stephen Downie, and Jin Ha Lee. "Stylistic Analysis on Reviews of Humanities Objects." Poster presented at the Chicago Colloquium on Digital Humanities and Computer Science, Nov. 2006, Chicago, Illinois. 2006a. <http://dhcs.uchicago.edu/abstracts/hu.pdf>

Hu, Xiao, J. Stephen Downie, and M. Cameron Jones. "Criticism Mining: Text Mining Experiments on Book, Movie and Music Reviews." *Digital Humanities 2006 Conference Abstracts.* Paris: CATI, Université Paris-Sorbonne, 2006b. 88-93.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis. "Text Genre Detection Using Common Word Frequencies." *Proceedings of 18th International Conference on Computational Linguistics, July 2000, Saarbrücken, Germany.* 2000.

What is Transcription?

Claus Huitfeldt *(Claus.Huitfeldt@fil.uib.no)*
University of Bergen
C. M. Sperberg-McQueen *(cmsmcq@acm.org)*
World Wide Web Consortium
MIT Computer Science and AI Laboratory

Background

One common task in digital humanities is the creation of digital representations of cultural artifacts, often in the form of transcriptions of existing physical documents. When we say, though, that a resource is a transcription of a particular document, what do we mean? What inferences are licensed by that claim? Given a transcription and adequate knowledge of the rules it follows, what do we learn about the document of which it is a transcription?

This paper proposes a formal account of transcription, as a way of elucidating the concept as it applies in scholarly editing and in the creation of digital resources. This approach may also provide insight into the nature of electronic representations of cultural artifacts, digital preservation, and text encoding more generally. It may have practical consequences for quality assurance and for the representation of markup semantics in formal systems.

In this abstract, we first present the basic outlines of our approach and our model of transcription, for simplicity's sake addressing only the very simplest cases. We then address some issues pertaining to less simple cases and discuss elaborations of the model. Because this is work in progress, the abstract is more sketchy on the latter issues, on which we expect to report more extensively in the paper as presented at the conference.

General outline of the approach

In general, one document (the transcription, T) is said to be a transcription of another document (the examplar, E) if T was copied out from E with the intent, successfully achieved, of providing a faithful representation of a text as witnessed in E. Often the purpose is to make this representation easier to make use of, than E itself. (For example, T may be easier to read or duplicate than E, or may be able to travel while E cannot.)

This description refers crucially to the intent with which T is made and the uses to which it may be put. These matters may be essential in deciding whether T is a transcription of E. In this paper, however, we investigate to what extent it is possible to characterize, without taking such matters into account, the relationship between T and E which allows the one to serve as a record of information derived from the other. In the simplest cases, where both E and T have an unproblematic, clearly legible inscription, this relation is simple: the two documents contain the same sequence of letters, spaces, punctuation marks and other symbols.

But what does it mean to say that a manuscript letter is "the same" as e.g. a typed or printed transcription of that manuscript? In simple cases, the question rarely arises, yet in transcriptions of historical documents it may be urgent. The difficulty of reading often consists precisely in the difficulty of deciding which letters the various marks on a manuscript page indicate (if any), and may be one reason for transcribing in the first place.

In order to establish a more precise account of the "sameness" of sequences of letters, we begin by distinguishing the concepts document, mark, token and type.

By a document we understand an individual physical object containing marks. A mark is a perceptible feature (normally something visible, e.g. a line in ink) of a document. Marks may be identified as tokens in so far as they are instances of types, and collections of marks may be identified as sequences of tokens in so far as they are instances of sequences of types. In other words, a mark is a token if, but only if, it is understood as instantiating a type.

The distinction among marks, tokens, and types may be applied at various levels: letters, words, sentences, texts. The identification and separation of the marks, tokens, and types of a document requires a competent reader. A theory of interpretation or of reading is not the concern of our paper, however — our goal is only to elucidate what it means to say that T is a transcription of E, however the interpretive processes resulting in T are properly understood.

Based on this outline of the relation between E and T and the distinction among documents, marks, tokens, and types, the following gives a first approximation to a more formal account of the meaning, in the simplest cases, of the statement that T is a transcription of E:

1. E and T are documents. (It follows that E and T each contain marks.)
2. The marks of each document are interpreted as instantiating a sequence of types:
    - Each mark is identified as constituting one and only one token.
    - Each token is identified as instantiating one and only one specific type.
    - To the tokens of each document, an ordering is assigned which is total, relative and linear.
3. The two sequences of tokens thus identified instantiate identical sequences of types.

A more explicit account of these points is best given not in prose but in a more formal notation. The full paper will provide translations of the description above into the modeling language Alloy [Jackson 2006]; for brevity, these formalisms are omitted from this abstract.

This simple account agrees well enough with common usage. Vander Meulen and Tanselle (1999), for example, describe transcription as "the effort to report — insofar as typography allows — precisely what the textual inscription of a manuscript consists of". We take "textual inscription" to denote the physical tokens of the document, and the method of reporting the textual inscription of the exemplar to be the creation of another document whose textual inscription produces (when rightly read) the same sequence of abstract types.

Even commonplace phenomena require further elaboration of this model (see below), but this simple model already illustrates some essential features of transcription.

- In a transcription, "an A is an A is an A" — one instance of a letter type is (for purposes of transcription) interchangeable with any other. Detailed information about the shape of individual marks is lost; the only salient information about the token in the exemplar is which type it maps to.
- The set of letter types distinguished in a transcription helps define that transcription; the choice of an appropriate inventory is a basic responsibility of the transcriber and a fertile source of disagreements.
- The formal relation between exemplar and transcript is symmetric and transitive. (Not necessarily so the concept of transcription — just the formal relation.)

Application and elaboration

A satisfactory formal account of transcription must deal with complications ignored above and will require elaboration of the basic outline. Our account has some affinity with Nelson Goodman's theory of notation [Goodman 1976]. We have no general characterization of these complications yet, so we introduce them here in the form of examples:

Some transcriptions must account both for systematic ambiguity of tokens in the writing system (a particular shape sometimes instantiates one type, sometimes another) and for uncertainty

about which token is actually present. These cases present special difficulties for the choice of the appropriate type inventory.

The mapping from mark to type may depend on context: marks of similar shape may be read now as commas, now as full stops.

Paleographic transcriptions often mark line breaks by vertical bars; the type sequence in the transcript thus appears to include items not present in that of the original. Shall the line breaks of the exemplar be understood as a special kind of letter token?

Some transcribers normalize or regularize the spelling of the exemplar, sometimes silently. The unit of transcription, in such cases, is the lexical item or word form, rather than the letter.

Scribal abbreviations pose a challenge: expanding them destroys the one-to-one correspondence between the tokens in the original and in the transcript. How can such expansions best be modeled? Perhaps each token in the exemplar corresponds to one or more tokens in the transcript? But then what of shorthand transcriptions made for private use, in which abbreviations are not expanded but introduced? The relation between the type sequences of exemplar and transcript must be more complex than simple identity.

Sometimes transcriptions include marks to which no mark in the exemplar seems to correspond. Vertical lines in a paleographic transcription may (as suggested above) be interpretable as transcribing line breaks in the exemplar. But if the transcription provides line numbers as well, it's difficult to stretch the notion of allographic variation to cover them. Similarly, transcriptions may contain expansions of abbrevations, supply letters or words assumed to be left out from the exemplar by slips of the pen, transcriber's comments on the wording or physical appearance of the examplar, etc., i.e. “additional” content not found in the exemplar.

The most fundamental problem, however, is given by the silent assumption, in the preceding remarks, that the text represented in a particular document, and thus the document itself, can usefully be represented as a one-dimensional sequence of types. This view appears painfully simplistic in the light of recent years' work on XML and text encoding. Even in conventional printed books, characters seldom have only one possible sequence. And digital representations seldom limit themselves to preserving the sequence of graphemes in the exemplar; normally, they use markup to identify larger textual units and make explicit aspects of the text which may lack any conventional orthographic representation.

One might take the structural units identified by markup (e.g. the elements of an XML document) as instantiating complex types; such an approach would in some ways resemble Goodman's analysis of ‘characters’, with the markup used to make the identification of complex characters explicit and reliable. Alternatively, one might treat marked up documents as representing a mixed notation, with character data content operating in one way, and markup in a different way. Goodman, for example, analyses drama as a mixed notation, in which the dialog functions as a “score”, but the stage directions function as a “script”; the analogy with markup may prove informative.

Whatever the approach, a key challenge for any formal treatment of transcription is to make it fit into the more powerful notions of textual structure reflected in modern markup systems. This is the most important topic to be addressed in the further development of the formal model of transcription.

Bibliography

Driscoll, Matthew. "Levels of Transcription." *Electronic Textual Editing.* Ed. Lou Burnard, Katherine O'Brien O'Keeffe and John Unsworth. New York: Modern Language Association of America, 2006. 254-261.

Goodman, Nelson. *Languages of Art.* Indianapolis, IN: Hackett Publishing Company, 1976.

Huitfeldt, Claus. "Multi-dimensional Texts in a One Dimensional Medium." *Computers and the Humanities* 28 (1995): 235-241.

Huitfeldt, Claus. "Philosophy Case Study." *Electronic Textual Editing.* Ed. Lou Burnard, Katherine O'Brien O'Keeffe and John Unsworth. New York: Modern Language Association of America, 2006. 181-196.

Jackson, Daniel. *Software Abstractions: Logic, Language, and Analysis.* Cambridge, MA: MIT Press, 2006.

Kline, Mary-Jo. *A Guide to Documentary Editing.* Second edition. Baltimore, MD: Johns Hopkins University Press, 1998.

Robinson, Peter. *The Transcription of Primary Textual Sources Using SGML.* Office for Humanities Communication Publications, Number 6. [Oxford: OHC], 2004.

Robinson, Peter, and Elizabeth Solopova. "Guidelines for the Transcription of the Manuscripts of the Wife of Bath's Prologue." *The Canterbury Tales Project Occasional Papers Volume I.* Ed. Norman Blake and Peter Robinson. Office for Humanities Communication Publications, Number 5. [Oxford: OHC], 1993.

Stevens, Michael E., and Steven B. Burg. Walnut Creek, CA: Altamira Press, 1997.

Vander Meulen, David L., and G. Thomas Tanselle. "A System of Manuscript Transcription." *Studies in Bibliography* 52 (1999): 201-212.

Un Outil pour un Nouveau Savoir Musical

Louis Jambou *(ljambou@paris4.sorbonne.fr)*
Lexique Musical de la Renaissance - Patrimoines Musicaux (EA 2560)
Université Paris-Sorbonne
Florence Le Priol *(flepriol@paris4.sorbonne.fr)*
LaLICC (FRE2919)
Université Paris-Sorbonne/CNRS

La musicologie a toujours été une discipline multi- et interdisciplinaire, de nombreuses disciplines étant pertinentes à l'étude de la musique et des textes musicaux: l'histoire, l'ethnologie, la sociologie, l'informatique, la linguistique... Le travail présenté ici illustre cette interdisciplinarité. En effet, notre outil fait appel à des compétences en histoire musicale et générale, théorie musicale, linguistique et informatique et fait collaborer des chercheurs du groupe«Lexique Musical de la Renaissance (LMR)» (EA 2560, Patrimoines Musicaux, Université Paris-Sorbonne) et du laboratoire«Langages Logiques, Informatique Cognition Communication (LaLICC)» (FRE2919, Université Paris-Sorbonne/CNRS).

Cet outil se présente sous la forme d'un dictionnaire du langage musical depuis la naissance de la théorisation musicale en langue vernaculaire jusqu'au seuil de la formation du langage tonal. Les entrées de ce dictionnaire sont fournies par les textes eux-me mes; textes essentiellement en langues néolatines à cause me me de la matrice de la théorisation musicale: le latin. Il réunit des citations se référant aux modalités d'élaboration, de production, de réception et circulation des savoirs théoriques musicaux qui constituent la base de données du projet. Son dictionnaire formera le fondement de la réflexion musicologique: réflexion unilingue définissant les champs sémantiques; réflexion interactive multilingue établissant des rapports entre les termes et les concepts.

Le but final de l'outil est sans aucun doute la musique, c'est-à-dire la partition produite au moment étudié. Mais au lieu de lui appliquer une théorisation a posteriori et des schémas musicaux/mentaux qui ne lui appartiennent pas, il est préférable d'aller d'abord aux textes théoriques et pratiques de la musique produite, ainsi qu'aux témoignages philosophiques, littéraires ou didactiques sur la discipline. L'étude de ceux-ci, leur élaboration à partir de leur contexte culturel et linguistique, fournira un outil propre à saisir et à comprendre nouvellement la musique de ce temps sans pour autant chercher à résoudre la dichotomie entre langage et musique.

Des dictionnaires techniques de la musique existent: ils présentent tous la caractéristique d'offrir une terminologie appréhendée et définie à partir du moment où leurs éléments sont conceptualisés et offerts à la consultation des spécialistes, musicologues ou musiciens. Une contemporanéité, d'aujourd'hui, qui offre à l'observateur un outil d'analyse anachronique, pour le passé. Dans cette perspective, la réflexion sur le terme musical est donc faite après sa mise en oeuvre dans l'histoire par le musicien, le théoricien ou le praticien, au travers d'un rideau temporel et de la superposition de strates-culturelles, techniques et musicales, multiples.

Depuis des décennies la musicologie a, en partie, fléchi ces orientations et le présent projet s'inscrit dans une lignée qui n'est pas totalement novatrice. Nous remarquerons d'abord que bien des études inscrivent leur réflexion sur le texte musical ancien dans le contexte de la production théorique contemporaine. Ainsi, pour s'en tenir à un exemple, les derniers travaux de Margarita Bent (« Diatonic ficta»; *Early Music History*, 4, 1984; trad. Français, Paris, Minerve,2003) ou Bonnie J. Blackburn (« On compositional Process in the Fifteenth Century»; *Journal of American Musicological Society*, 1987, trad. Français, ibid.) sont en effet proches de la pensée musicale théorisée de l'époque avant que de l'appliquer à leur approche analytique des oeuvres mises à l'épreuve. Mais les critères de sélection de ces textes ne sont jamais justifiés et ce ne sont que quelques textes choisis qui viennent appuyer ou infléchir la lecture dont l'on notera qu'elle se fait toujours dans un rapport triangulaire (texte théorique + musicologue auteur + collègue musicologue pris à témoin dans l'acquiescement ou la réfutation de l'argumentation). Notre travail s'inscrit dans la lecture textuelle de l'époque mais celle-ci se veut sinon exhaustive, du moins totalisante. Par ailleurs nous ne chercherons pas a priori à adopter une attitude favorable ou contraire à l'une ou l'autre orientation musicologique actuelle. Seuls les résultats futurs permettront, ou non, d'adopter une nouvelle épistémologie des oeuvres musicales de la période considérée.

Durant la seconde moitié du XXe siècle, d'autres dictionnaires tels que le *« Handwörterbuch der musikalischen Terminologie (Htm) »* fondé par Hans Heinrich Eggebrecht ouvrent également une voie nouvelle d'approche de la lexicologie musicale. Cependant sa remarquable entreprise reste marquée par les prospectives antérieurs et ne s'en affranchit que dans les années 1970-1980. A partir de ces dates naissent, avec les nouvelles technologies, de nouveaux outils qui ont mis en ligne les textes originaux (Université d'Indiana, USA: textes en latin, français et italien) ou cherchent à en dresser des études lexicales (sur le latin: prof. Bernhard, université de Munich en collaboration de

l'université de d'Indiana; sur le corpus italien: prof. Wiering, université d'Utrecht).

Le présent projet est plus proche du«Lessico italiano del canto» (*Canto*) par son recours systématique aux textes contemporains et aux disciplines conjuguées de la linguistique et de la musicologie. Fondé en 1987 par le prof. Sergio Durante (université de Padoue), le programme *Canto* recherche la mise en rapport entre textes théoriques et musique mais cette recherche s'effectue dans une aire géographique déterminée et une temporalité définie. Cette recherche embrasse, en Italie, les paramètres du phénomène du chant de 1600 à 1800.

Notre étude vise à d'établir la genèse d'une pensée proprement lexico-musicale et de l'amplifier à d'autres aires linguistiques et à d'autres temporalités. L'originalité de notre travail consiste en la recherche d'une totale autonomie de la pensée musicale, tant dans la formation du thesaurus lexical et des citations que dans la réflexion menée sur le corpus.

Les objectifs finaux sont:

- de relire et repenser les traités de théorie et pratique musicales, et autres écrits véhiculant le langage sur la musique, de la période considérée;
- de dresser un corpus plurilingue des termes techniques, dans un sens générique, touchant la musique en ses moments de préparation, de production, de perception et de réception;
- de former un dictionnaire raisonné comprenant un répertoire de citations pertinentes des différentes langues techniques musicales vernaculaires en pleine élaboration dans leur saisie du phénomène sonore;
- de définir le niveau de compréhension du phénomène sonore en chaque langue et en des périodes distinctes c'est-à-dire de réfléchir aux vitesses d'implantation, de consolidation et d'évacuation ou permanence des termes/concepts dans la langue technique;
- de créer une nouvelle articulation entre les deux niveaux de langage en un projet sémiotique entre le langage théorique musical et la propre création de l'objet musical;
- de redéfinir ainsi un outil de lecture interprétative et d'analyse de la musique à partir de la perception des acteurs de l'époque par la création de nouveaux champs de contenus musicaux;
- de mettre à la disposition des chercheurs l'acquisition de ce nouveau savoir, par l'accès à la base de données.

Les étapes préparatoires au projet ont fait l'objet d'échanges entre les chercheurs afin de permettre de former des outils favorisant un travail pouvant e tre mené en commun mais également de façon autonome.

Le corpus bibliographique des langues principales (français, italien, espagnol) a été constitué. D'autres langues sont appelées à s'y joindre et demandent une approche bibliographique encore en cours de réalisation (portugais, catalan, roumain...). Le choix des textes, imprimés ou manuscrits, est établi au plus près de l'édition originale qui sert d'unique référence. Si celle-ci est inaccessible, il y aura lieu de recourir aux facsimilés ou films des originaux. A défaut l'édition critique la plus fiable est exploitée.

Un prototype de la base de données a été élaboré[1] avec le logiciel Microsoft Access. Gérée sur un poste unique, cette base était alimentée par les chercheurs de manière indirecte et n'était pas accessible en consultation. Chaque chercheur effectuait sa propre campagne de relevés de citations et la stockait dans un fichier Microsoft Word dont l'organisation avait été préalablement discutée. Les fichiers étaient envoyées à l'administrateur de la base de données qui y intégrait ces données.

Pour le développement du projet, deux constats se sont imposés:

- les chercheurs impliqués dans la constitution des ressources devaient pouvoir directement saisir leurs données et y accéder;
- toutes les ressources devaient e tre mise à la disposition de tous les musicologues, musiciens... par le biais d'internet.

Ainsi, du prototype développé avec un système de base de données fermé, nous sommes passé à un système ouvert, basé sur la trilogie Apache-PHP-MySQL, accessible sur internet (<http://www.pm.paris4.sorbonne.fr/LMR>).

La base de données fournit aujourd'hui un réservoir de citations et des définitions d'acceptions multiples à tout chercheur. Sa mise en ligne permet un questionnement élaboré par la création de liens de consultations ou d'interrogations qui intègrent les exemples musicaux et les illustrations iconographiques contenus dans les traités. Elle donne aux chercheurs participant à l'élaboration des ressources constitutives de l'outil, un accès sécurisé, pour la saisie, les modifications et les corrections de leurs données et l'organisation sémantique du dictionnaire.

Cet outil vise un nouveau savoir musical où la formulation d'une pensée musicale doit e tre linguistiquement autonome dans une aire géographique donnée. Il doit favoriser l'élaboration d'études spécialisées sur des champs précis de la théorie musicale et permettre de faire apparai tre des connexions entre ce corpus musical et la contemporanéité de l'état social, politique et le savoir culturel: pouvoir et institutions, philosophie et humanisme, nouvelles orientations scientifiques.

---

1. Réalisé par Alexandre Dutra-Cançado (gorupe LMR, Patrimoines Musicaux, Université Paris-Sorbonne), grâce à la collaboration

de Detlev Schumacher (groupe Canto, Bibliothèque Nationale de Florence)

Références

Musica e storia , X/1, Venise: Fondazione Ugo e Olga Levi, 2002. Actes du colloque tenu à Venise à la Fondation Levi, *La musica fra suono e parola: ricerche sul lessico musicale in Europa* (26-28 ottobre 2000 et corrodonnée par Fiamma Nicolodi et Sergio Durante. Communications du groupe LMR:

- Chevalier, Jean-Claude, Delport, Marie-France, "De la citation à la définition. Tañer et tocar. Ou: «jouer n'est pas souffler»", pp 15-26.
- Robledo, Luis, "El léxico musical en el contexto humanista español: la prosa didáctica y la preceptiva retórica",pp 151-164.
- Jambou, Louis, Dutra-Cançad, Alexandre, "Les genres grecs dans la théorie musicale de la Renaissance en langue vernaculaire: l'exemple de l'espagnol", pp 165-184.

De la lexicologie à la théorie et à la pratique musicales , actes du colloque tenu en Sorbonne le 16 juin 2001, textes réunis par Louis Jambou, Paris, Editions hispaniques, 2002. Contient les textes suivants:

- Durante, Sergio, «Per un incontro di lessicologia musicale», pp 7-10.
- Jambou, Louis, «La temporalisation dans les traités musicaux: l'exemple de Bermudo (1555) et de Sancta Maria (1565)», pp 11-22.
- Robledo Estaire, Luis, «Una taxonomi a ética de la mu sica: el Libro primero del espejo del pri ncipe christiano (Francisco de Monzón), pp 23-30.
- Olmos, Ángel Manuel, «L'acoustique chez Francisco Tovar et les théoriciens du début du XVIe siècle: le tapage des sphères», pp 31-41.
- Grasso Caprioli, Leonella, «Note per uno studio lessicografico sul canto in Rossini», pp 43-56.
- Chevalier, Jean-Claude, Delport, Marie-France,«De l'enseignement d'un corpus», pp 57-69.

Journée d'études autour des projets "LMR" (Paris-Sorbonne) et "Canto" (Univ. Padoue) , 1 avril 2000. Sorbonne. Avec la participation de Madame Grasso Caprioli (Un. de Padoue/Bologne) et Monsieur Schumacher (BN Florence).

Journée d'études «Musique ancienne en Sorbonne» , 2 juin 2006, Maison de la Recherche, Sorbonne, Paris. Présentation par Florence le Priol et Louis Jambou des Problématiques et enjeux du«Lexique Musical de la Renaissance» (LMR)

Digital Visualization as a Scholarly Activity

Martyn Jessop *(martyn.jessop@kcl.ac.uk)*

Centre for Computing in the Humanities
King's College London

Introduction

The numerous visual metaphors that describe cognitive processes hint at the nexus of relationships between what we see and what we think. We say we 'see' when we mean we understand, we try to organize and make our ideas 'clear' by bringing them into 'focus', and so on. When faced with tasks that require substantial thought or organization of ideas we will often reach for a pen and paper to 'sketch out' (another visual metaphor) our thoughts. We have a deep understanding that we can enhance our thought processes by finding ways of linking external perception with our interior mental processes. Graphic aids to thinking are not new but the development of computers has provided a new medium with remarkable functionality. This in turn offers the potential for new research methodologies that amplify cognition. These tools serve two distinct purposes. One of these is often described by the hackneyed phrase "A picture is worth ten thousand words".[1] This misses the true point of visualization as what is being described here is just a matter of transmission, of having high bandwidth to transmit large volumes of information. Of far greater importance is the ability of these tools to allow visual perception to be used in the creation or discovery of new knowledge. Knowledge is not transferred, revealed, or perceived, but is created through a dynamic process. This raises epistemological issues concerning visualization and points the way to an intellectual approach to the subject.

Computer visualization techniques began as a methodology for understanding the meaning of large volumes of numeric data. Scientists needed a means of visualizing the flood of data that can be collected by modern monitoring and measuring instruments. The National Science Foundation initiative on Scientific Visualization launched in 1985 led in a very short time to Scientific Visualization becoming recognised as not just a methodology but a discipline in its own right. A similar trend may now be developing in the Humanities. For example the London Charter seeks to establish principles for the use of 3D visualization in research and communication of cultural heritage that ensure the intellectual integrity of the methods

and outcomes derived from it. How is visualization being used in the humanities at the moment? Is there a potential counterpart to Scientific Visualization in the digital humanities – a field of 'humanistic visualization'? What issues does it raise? Where is the common ground and what are the intellectual issues involved?

Visualization in the Digital Humanities

There are many ways of structuring an examination of the use of visualization in the humanities; by discipline, by type of information structure and so on. To set the context for this paper I have chosen to look at the type of data that is being visualized. The boundaries of data types are sometimes blurred but a starting point could be as follows

- Numbers. Quantitative analysis and visualization has been an established tool in many humanities disciplines for a long time. It is found in generic statistical analysis software or embedded in specialised applications such as text analysis. It is gradually permeating into new areas of the humanities through work such as that of Franco Moretti (2005) who has argued for a wider application of quantitative methods in areas such as literary history.
- Text: Visualization techniques using tables and graphs have been commonplace in text analysis for many years. These visualizations are sometimes variants of statistical visualizations of numeric data (as in word frequencies) but in other cases they are specialised visual forms of text analysis. Projects such as the TAPoR and NORA Projects are developing imaginative new visual forms and applications. These new methods apply not just to the visual representation of the results of analysis but also to the visualization of the texts themselves. This aims to support interpretive scholarship by allowing areas or relationships of interest to be identified within large volumes of text. Projects have explored specific texts in this way, for example Dante's Inferno , Hume's Dialogues and The Shape of Shakespeare but there is substantial scope for a tool that could be applied to any text. The use of the word 'tool' here should not be taken to imply that this is a computational problem, this is far from the truth, as what is being grappled with here are conceptual problems. For example, what is a text? How should it be displayed visually? Humanities computing provides a medium with a myriad of possibilities of representation; uni-dimensional or two dimensional physical objects, abstract objects showing relations among words or between words and annotations, and animations. Further questions arise from this work; is there a need for representational as well as interpretive markup? What are the relationships between text visualization and the interpretation of texts? The visualizing of texts is also an area which links humanities computing work to the arts, for example the interactive installation Text Rain by Camille Utterback and Romy Achituv.
- Narratives and relationships. These can be grouped and referred to as diagrams . Edward Tufte has drawn public attention to this style of diagram; famous examples include the work of Playfair and Charles Minard's narrative graphic of Napoleon's ill-fated campaign against Russia in 1812. In many respects this has been a neglected field since the advent of digital tools because diagrams of this type are not easily implemented in software. They are also potentially of enormous value to the humanities as narratives and the study of relationships between people, events and artefacts are studied by many disciplines.
- Space. The study of spatial relationships and a sense of place occur in many humanities disciplines. This area is dominated by Geographical Information System (GIS) software but this was developed for scientific data and is not ideally suited to the qualitative data used in the humanities (Jessop, 2006). Digital dynamic maps are one of many alternatives offering media that are better suited to humanists (Jessop, 2006). The Electronic Cultural Atlas Initiative (ECAI) with its utilization of Timemap provides an indication of possible future developments.
- Time. Digital visualization provides a very powerful medium for temporal visualisation. Tools such as timelines allow one to explore the development of complex historical events and the inter-relationships between precursor events. They are of obvious value in the study of history where later events build upon earlier ones but they can also be applied elsewhere. Historiography and literary criticism are both histories of accumulated comments on a subject. Matt Jensen (2006) has developed a number of timeline tools which are intended to answer the styles of questions that are asked by humanists, for example in the case of political scandals questions of the style 'who knew what and when' or for exploring the response to an author's writing over a period of many years
- 3D Visualization. Much of this work has centred upon visualizations of the built environment. It is of interest to not only historians and archaeologists but also anyone who seeks to find out how the buildings of the past worked in human terms, for example the Pompey Theatre and Theatron projects are based on historical and archaeological data but are primarily of interest to scholars of theatre studies. 3D visualization is of especial interest in the context of this paper because there is currently great deal of work focused on defining not only good practice (see ICT Methods Network) but also principles for maintaining the intellectual integrity of such work, for example the London Charter. It may therefore provide pointers for similar work in the

development of a broader humanistic visualization as a whole.

Any demarcations between the applications of visualization between different disciplines are misconceived. There is much common ground offering considerable potential for humanities computing and the digital humanities. This is where we need to focus our attention if digital visualization is to achieve recognition as a rigorous intellectual activity in research and teaching.

Conclusion

We accept that texts and documents are produced as readings resulting from acts of interpretation between the reader and the text; we now need to regard images in the same way. Every representation, visual or otherwise, is an effort to structure an argument and as such it is a rhetorical device. We need to understand the relationship between what is being communicated and how it is being communicated.

Information graphics, and indeed images generally, can be considered as historical artefacts themselves, filled with interesting incidental and substantive information embodied in their production, style, and graphical properties. But perhaps more importantly, they are expressions of procedures for generating knowledge through the act of visualization and ways of displaying knowledge embodied in visual imagery (Drucker, 2003).

Visualization addresses epistemological and pedagogical issues that are common to the digital humanities and are at the forefront of the developing discipline of humanities computing. This is a vast topic the main aim here is to identify some of the underlying intellectual issues arising from visualization and the use of images in digital humanities scholarship.

---

1. This is commonly believed to be based on a “Chinese proverb” however Paul Martin Lester believes that it was in fact made up by the advert writer Federick R. Barnard. See <http://www5.Fullerton.edu/les/ad.html> and Printers’ Ink, March 10, 1927.

Bibliography

Barthes, Roland. *Elements of Semiology.* New York: Hill and Wang, 1968.

Bertin, Jacques. *The Semiology of Graphics: Diagrams, Networks and Maps.* Trans. William J. Berg. 1967. Madison, WI: University of Wisconsin Press, 1983.

Card, Stuart K., Jock Mackinlay, and Ben Shneiderman, eds. *Readings in Information Visualization: Using Vision to Think.* San Diego, CA: Academic Press, 1999.

Drucker, Johanna. "Graphesis: Visual Knowledge Production and Representation ." 2003. Accessed 2006-10-20. <http://www.noraproject.org/reading.php>

Gombrich, E. H. *Art and Illusion.* New York: Bollingen, 1961.

Ivins, William M., Jr. *Prints and Visual Communication.* Cambridge, MA: MIT Press, 1969.

Jessop, Martyn. "The Visualization of Spatial Data in the Humanities." *Literary & Linguistic Computing* 19.3 (2004): 335-350.

Jessop, Martyn. "Dynamic Maps in Humanities Computing." *Human IT* 8.3 (2006): 68–82. <http://www.hb.se/bhs/ith/3-8/mj.pdf>

Keeler, Mary. "The Place of Images in a World of Text." *Computers and the Humanities* 36.1 (2002): 75-93.

McCarty, Willard. *Humanities Computing.* Basingstoke, UK: Palgrave Macmillan, 2005.

McCormick, B. H., and T. H. DeFanti. "Visualization in Scientific Computing." *Computer Graphics* 21.6 (1987).

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary Theory.* London: Verso, 2005.

Norman, Donald A. *Things that Make Us Smart.* Reading, MA: Addison-Wesley, 1993.

Rockwell, Geoffrey. "What is Text Analysis, Really?" *Literary & Linguistic Computing* 18.2 (2003): 209-219.

Rohrer, Randall, David Ebert, and John Sibert. "The Shape of Shakespeare: Visualizing Text Using Implicit Surfaces." *Proceedings of the IEEE Symposium on Information Visualization.* 1998.

Scaife, Mike, and Yvonne Rogers. "External Cognition: How do Graphical Representations Work?" *International Journal of Human-Computer Studies* 45.2 (1996): 185-213.

Schreibman, Susan. "Computer-mediated Texts and Textuality Theory and Practice." *Computers and the Humanities* 36.3 (2002): 283-93.

Sinclair, Stéfan. "Computer Assisted Reading: Reconceiving Text Analysis." *Literary & Linguistic Computing* 18.2 (2003): 175-184.

Tufte, Edward R. *The Visual Display of Quantitative Information.* Cheshire, CT: Graphic Press, 1983.

Unsworth, John. "Forms of Attention: Digital Humanities Beyond Representation." Paper presented at "The Face of Text: Computer-Assisted Text Analysis in the Humanities," the third conference of the Canadian Symposium on Text Analysis

(CaSTA), McMaster University, November 19-21, 2004. 2004. Accessed 2006-10-20. <http://www3.isrl.uiuc.edu/%7Eunsworth/FOA/>

Unsworth, John. "New Research Methods for the Humanities." The Lyman Award Lecture, National Humanities Center, November 11, 2005. 2005. Accessed 2006-10-20. <http://www3.isrl.uiuc.edu/~unsworth/lyman.htm>

Wurman, Richard Saul. *Information Design.* Cambridge, MA: MIT Press, 1999.

Web Resources

- A Visualization of Dante's Inferno <http://urizen.village.virginia.edu/hell/>
- Electronic Cultural Atlas Initiative (ECAI) <http://ecai.org/>
- ICT Methods Network <http://www.methodsnetwork.ac.uk/index.html>
- King's Visualization Lab <http://www.kvl.cch.kcl.ac.uk/>
- The London Charter <http://public-repository.epoch-net.org/TheLondonCharter_v1.pdf>
- Nora Project <http://www.noraproject.org/description.php>
- Pompey Theatre Project <http://www.kvl.cch.kcl.ac.uk/pompey.htm>
- Text Analysis Portal for Research (TAPoR) Project <http://www.tapor.ca/>
- TimeMap <http://www.timemap.net/>
- Theatron <http://www.theatron.org/>

The Other Side of the Rug: TokenX on the *Willa Cather Archive*

Andrew Wade Jewell *(ajewel2@unl.edu)*
University of Nebraska-Lincoln
Brian L. Pytlik Zillig *(bpytlikz@unlnotes.unl.edu)*
University of Nebraska-Lincoln

> [Schools] can only teach those patterns which have proved successful. If one is going to do new business the patterns cannot help, though one does not deliberately go out to do that. My Ántonia, for instance, is just the other side of the rug, the pattern that is supposed not to count in a story.
>
> (Willa Cather, 1925)

Tools for text analysis have long been a focus of many digital humanities scholars, yet the results produced by those tools are rarely utilized in typical literary and cultural criticism. Though the reasons for this disconnect are varied, we believe two primary hurdles are visibility and accessibility. Specifically, most text analysis tools and research are not found on the sites—like thematic digital text archives—that scholars use most, and most text analysis tools are not designed for ordinary scholars, but are meant for those with more technical sophistication.

Our essay discusses a novel approach to filling this gap between scholars and text analysis, an ongoing collaborative experiment in humanities computing to bring a "new business" that willfully counters the "patterns which have proved successful" in literary criticism in order to add a new dimension to literary research. In "The Other Side of the Rug: TokenX on the *Will Cather Archive*," we discuss the application of Brian Pytlik Zillig's text analysis, visualization, and play tool, TokenX (<http://tokenx.unl.edu>) to the *Willa Cather Archive* (<http://cather.unl.edu>), a free, educational resource dedicated to the study of Willa Cather's life and writings and edited by Andrew Jewell. When this application of TokenX debuts in the summer of 2007, scholars will be able to analyze the entire corpus of Cather's fiction, from her first college publications to her final short story. In many ways, our project is in the tradition of research that seeks to use text analysis tools to arrive at, in Ramsay and Steger's language, "suggestive patterns" to "enable critical reflection in literary study." However, in creating this tool for application on the *Cather Archive*, we were faced with two distinct challenges: (1) how to develop

TokenX to make it capable of the envisioned crossdocument analysis which is sensitive to changes over time, and (2) how to design an interface that would make sophisticated text analysis a manageable, useful tool for the widest possible range of Cather scholars, scholars unaccustomed to using such tools.

The project is a interdisciplinary collaboration between Pytlik Zillig, a Digital Initiatives Librarian with specialization in XSLT and text analysis, and Jewell, Assistant Professor of Digital Projects with a Ph.D. in American literature, both of the University of Nebraska-Lincoln's Center for Digital Research in the Humanities. The paper, likewise, is a collaborative work that represents two distinct but complementary perspectives on the issue.

Pytlik Zillig asserts that, for digital humanities, the development and ready availability of tools to assist in the noticing and appreciation of texts is of increasing importance. Unsworth has observed that “by paying attention to an object of interest, we can explore it, find new dimensions within it, notice things about it that have never been noticed before, and increase its value." For the first iteration of the TokenX/Cather collaboration, the TokenX tool has generated a word frequency data set containing nearly half a million data cells. These data reveal the frequency of usage of words in fifteen TEI-encoded XML texts, representing Cather’s complete corpus of book-length fiction. This data set will be available for dynamic, user-centered queries to assist in formulating theories and facilitating explorations of Cather’s changing diction over time. (See Figure 1 for initial results on ten sample terms within Cather's corpus, detailing her usage of certain "body" words within each of her published books of fiction.) If a text’s common words are, as John Burrows suggests, "a barely visible web that gives shape to whatever is being said" (323), then it must be the ambition of tools such as TokenX to expose the dimensions of that web for further inquiry. (Plans are underway to add TokenX to the Text Analysis Portal for Research [TAPoR], which will enable TokenX users to visualize, analyze and play with documents stored on the TAPoR portal.)

Jewell argues that introduction of and experimentation with new tools for engagement with literary texts are an important way to make author-centric sites, like the *Willa Cather Archive*, models of innovation. The audience of the *Cather Archive* is not one inherently inclined to think about literary texts numerically. The onus on the designer, then, is to consider the sorts of research questions driving Cather and American literature scholarship and to make this tool something that would contribute to tackling such questions. For example, how might text analysis contribute to a scholar exploring Cather's work with a cultural studies or gender studies approach? What about a scholar looking at contexts, themes, or references within a single novel, or the fiction of a finite span in Cather's career? How might this tool aid a researcher interested in tracking evolutions in Cather's prose over a long period of time? (See Lindemann and the program of the 2005 International Cather Seminar for examples of these approaches). By designing an interface that is sensitive to a range of scholarly inquiry, one that allows for a significant amount of flexibility and user input, TokenX on the *Cather Archive* can represent an innovative use of digital research that is brought into the mainstream of scholarship. By demonstrating sensitivity to researcher interests and seeking to design broadly useful tools, we demonstrate to colleagues that such tools are not just for specialists, but can enrich diverse arguments, any that find foundation— as most literary scholarship does—on the use of words.

Bibliography

Burrows, John. "Textual Analysis." *A Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing, Ltd, . 323-347.

Lindemann, Marilee. *A Cambridge Companion to Willa Cather*. Cambridge: Cambridge University Press, 2005.

Program of the International Cather Seminar 2005. . <http://cather.unl.edu/seminar2005/sem_schedule.html>

Ramsay, Stephen, and Sara Steger. "Distinguished Speakers: Keyword Extraction and Critical Analysis with Virginia Woolf’s The Waves." Paper presented at Digital Humanities 2006, Paris Sorbonne, 5-9 July 2006. 2006. <http://www.allc-ach2006.colloques.parissorbonne.fr/DHs.pdf>

Unsworth, John. "Forms of Attention: Digital Humanities Beyond Representation.." Paper presented at The Face of Text: Computer-Assisted Text Analysis in the Humanities, the third conference of the Canadian Symposium on Text Analysis (CaSTA), McMaster University, November 19-21, 2004. 2004. <http://www3.isrl.uiuc.edu/~unsworth/FOA>

Macro Analysis (2.0)

Matthew Jockers *(mjockers@stanford.edu)*
Stanford

At the 2005 meeting in Victoria, I presented a paper with the revised title, "A Macro-Economic Model for Literary Research." That paper marked an early phase in a project aimed at leveraging large, digital, text corpora for what I now call "Macro-Analysis." The work presented in Victoria was largely a proof-of-concept using an electronic bibliography of 775 works of Irish-American literature. For that paper I performed a dumbed down sort of macro-analytic text analysis using metadata fields and the titles of works in the bibliography. Today the software has improved significantly, and in place of the title analysis of 2005, this paper offers results derived from a macro-analysis of the full text of 1125 British and American novels from the 19th century. In the presentation, I provide a general overview of the tool(s) and an explanation of the methodology employed in the analysis.

The tools and techniques I have develop utilize both supervised and unsupervised text-mining techniques. The supervised techniques allow for a focused analysis in which a researcher probes the corpus for items meeting a specific research criteria. A very simple example might involve tracing the behavior or "frequency" of some "signal" (linguistic pattern, literary theme, or author style) over the course of the corpus. I should note here that while this process sounds similar in some ways to the supervised machine learning approach being used by the NORA project, it is specifically not like NORA in that I am not employing machine learning or utilizing previous identified, "marked," training data. Instead the "signal" is developed ad hoc by the researcher/user.

An easy way to understand the project is to visualize the tool being used: A user is given an interface that allows for the usual sort of corpus searching. The user performs a corpus wide search for some term (or other feature such as a word cluster or syntactical pattern). The result page reports all of the texts in the corpus in which the search term(s) is found, and then the user is given a "toolbox" of macro-analytic tools with which to process and analyze the result set. These tools are varied and perform diverse sorts of analysis.

A topic-modeling tool, for example, provides the ability to harvest the salient themes from the text in the result set. The user is thus able to say, for example, that works in the corpus that contain word "x" show a predominance of the "n" topic. In my own work with ethnic American literature, I have found this tool valuable in assessing and quantifying the dominant themes that occur in works where ethnic markers (words denoting race or ethnicity) occur. This technique is derived from the work of David Newman and his research team at the University of California—Irvine.[1]

Another tool offers a type of literary time series analysis. Figure 1 shows a graph produced by my "timeline" tool.

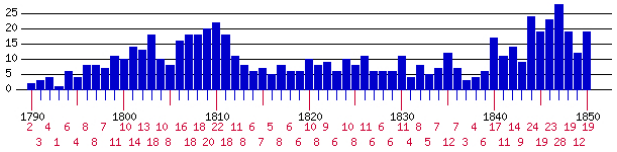


Figure 1

With the timeline tool, the results of any query of the corpus can be mapped over time. Say, for example, that I am interested in how some textual feature evolves over time. I perform the necessary query to isolate the occurrences of that feature and then choose the timeline tool. The resulting graph provides a visual time-series analysis of the frequency of the pattern. The graph shown here was produced after a simple search for occurrences of the word "romance" in the titles of 7300 novels from the 18th and 19th century. The raw counts are displayed in red beneath each year. In addition to providing this timeline of raw hits (figure 1), a second graph (figure 2) is also produced that shows a normalized result in the form of "hits-per-100" texts.

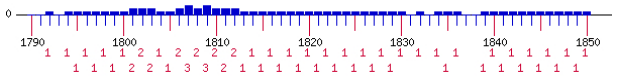


Figure 2

In this case, the normalized graph is particularly revealing because it shows that the frequency of "romance" as a key word in titles is not especially noteworthy. Aside from a very brief period (1800-1810) where the word appears in 2-3% of all titles, its use is steady at 1% or 1 occurrence per 100 titles in a given year.

The macro-analytic tools developed in this research exist as both command line applications and as a (beta) extension to the open-source eXtensible Text Framework (XTF) application developed by Martin Haye and the California Digital Library.[2] The successful implementation of the tools into XTF has been achieved with the assistance of Stanford Undergraduate digital humanities major Jenny Loomis who will spend the final five minutes of this presentation giving a live demonstration of the tool as implemented into the XTF application.

1. Newman's work is profiled on line at the following sites: <http://arstechnica.com/news.ars/post/20060802-7408.html> <http://www.ics.uci.edu/community/news/press/view_press?id=51>
2. See <http://www.cdlib.org/inside/projects/xtf/>

Preserving Information About Linearization in Document Graphs

Lars G. Johnsen *(lars.johnsen@lili.uib.no)*
University of Bergen
Claus Huitfeldt *(Claus.Huitfeldt@fil.uib.no)*
University of Bergen

Many operations are more conveniently performed on a graph representation than on a linear representation of a marked up document, and vice versa. Therefore, it is sometimes important to ensure that no relevant aspect of the information contained in a document represented in one of these forms is lost or distorted when the document is converted to the other form.

Conventional methods for converting between XML documents and their graph representations [W3C 2000] are typically seen to preserve such information; standards and methods have been established for ensuring what is in most contexts considered full preservation of all relevant aspects of the linearization [W3C 2001].

However, what is considered relevant may of course vary, depending on context of use. It would probably be hard to find serious arguments to the effect that literally all aspects of the linear representation of a document are relevant for any generally interesting use. Typically, conventional conversion methods are not guaranteed to preserve e.g. attribute order, declaration order, and insignificant whitespace. But it is not hard to find complaints about, for example, lack of preservation of attribute order in certain applications.

Our focus in this paper is on methods for the preservation of element serialization order in marked up documents which make use of mechanisms for representing non- hierarchic complex structures such as overlapping, discontinuous and virtual elements. (For convenience, we use the term "complex structures" to refer to such phenomena.) We do not wish to claim that preservation of element order is always or even generally relevant, our aim is limited to providing a method for such preservation in cases where it is considered relevant.

The customary graph representation of XML is in the form of an "XML tree", a restricted kind of directed acyclic graph (DAG). More specifically, XML trees are DAGs with single parenthood and total ordering on leaf nodes. For certain

purposes, however, a different kind of graph representation has been proposed, the so-called Goddag [Sperberg-McQueen and Huitfeldt 2000]: Roughly, Goddags are like XML trees except that they allow multiple parenthood and do not require a total ordering on leaf nodes; leaf nodes may be ordered only relative to their immediate parents. (Thus, XML trees constitute a subset of Goddags.)

For certain purposes this data structure provides a more convenient representation of complex structures than XML trees. Documents using different XML mechanisms for representing such structures in linear form (e.g. milestones, fragmentation, virtual elements etc. [Barnard et.al. 1995, Sperberg-McQueen and Huitfeldt 1999]) can be mapped on to Goddags, though not without knowledge of application-specific semantics of the markup vocabulary. The experimental markup system TexMecs [Sperberg-McQueen and Huitfeldt 2001] offers mechanisms for the representation of complex structures which can be mapped on to Goddags independently of such knowledge.

However, in both cases, i.e. whether the graph is built from XML or TexMecs, reserialization from the graph is not in general guaranteed possible without changes to the structure and order of elements in the original linearization. For example, if an XML document has used milestones or fragmentation of elements to represent overlapping elements it is possible to build a Goddag representing the non-hierarchic structure of the document. But when reserializing back to XML, the Goddag does not contain any information about which elements to represent as milestones or as fragmented elements.

Similarly with TexMecs: Some element structures can be represented by alternative serialization constructs, and the Goddag as currently defined does not preserve information about the choice of construct in each particular case. In TexMecs the problem is made more severe by the fact that the graph does not, in the case of e.g. virtual or discontinuous elements, preserve complete information about the serial orderof elements in the original input.

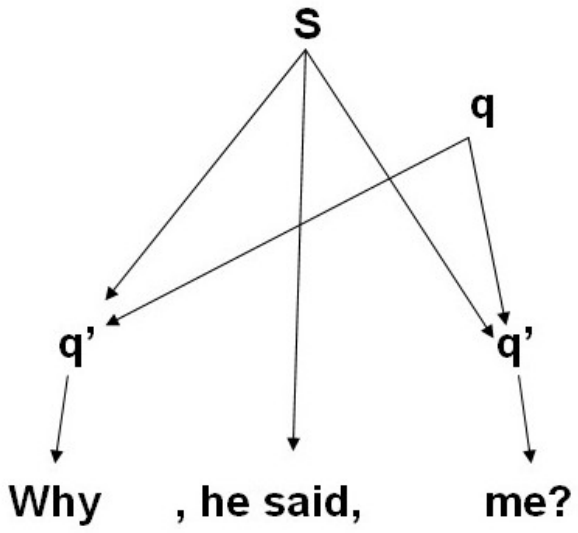
Consider the following example, marked up according to TexMECS, and illustrating how a discontinuous constituent element `<q>` may be represented.

(1) `<s|<q|Why|-q>, he said,<+q| me?|q>|s>`

In TEI-based XML, the example could e.g. be marked up as:

(1') `<s><q part="I">Why<q/>, he said,<q part="F"> me?</q></s>`

The resulting Goddag, whether based on the XML or the TexMecs input, would normally look like this:[1]



Since the second leaf node (containing the string ", he said,") does not share any parent with either of the two other leaf nodes, it is not ordered with respect to these. Therefore, the linearization in (1) is equivalent to the following two linearizations, both placing the second leaf node in a position relative to the other two that it does not have in (1):

(3) `<s|<q|Why me?|q>, he said,|s>`
(4) `<s|, he said,<q|Why me?|q>|s>`

Thus, the Goddag in (2) would be the same whether built from (1), (3) or (4). Similarly, a choice whether to linearize (2) in the form of (1), (3) or (4) will either have to be arbitrary, or based on purely practical considerations.

A solution to the linearization problem lies, we propose, in considering the Goddag used for representing marked up text as a path ordered directed acyclic graph; a Podagra. Building a Podagra from (1), we get a graph consisting of three paths, in the order indicated as follows:[2]

(5)
1. s → q → "Why"
2. s → ", he said,"
3. s → q → " me?"

Building a Podagra from (3), however, produces the following path order:

(6)
1. s → q → "Why me?"
2. s → ", he said,"

whereas from (4) we get the following paths:

(7)
1. s → ", he said,"

2. s → q → "Why me?"

The Podagras (5), (6) and (7) all correspond to the Goddag (2), but each maps uniquely to the linearizations (1), (3) and (4), respectively.

In the full paper, we will present an algorithm yielding Podagras from TexMECS documents containing different linearizations also of overlapping and virtual elements. We intend to show how path ordered Goddags can faithfully restore the original linearization of such documents.

Since TexMECS is a purely experimental markup language, these results may be said to have limited practical relevance. However, a number of projects currently build Goddags from XML-encoded documents (by the use of application-specific semantics). Therefore, we also hope to indicate how the proposed method may used for preservation of the original linearization of XML documents using well-known methods for representation of overlapping, discontinuous and virtual elements.

---

1. For simplification, we are consciously ignoring certain unresolved issues concerning the representation of discontinuous elements in Goddags [Huitfeldt and Sperberg-McQueen 2006].
2. The simplicity of the example allows us to indicate nodes by their generic identifiers. I.e. the three occurrences of "s" all indicate the single node labelled s, and the two occurrences of "q" indicate the single node labelled "q".

Bibliography

Barnard, David, Lou Burnard, Jean-Pierre Gaspart, , C. Michael Sperberg-McQueen, and Giovanni Battista Varile. "Hierarchical Encoding of Text: Technical Problems and SGML Solutions." *The Text Encoding Initiative: Background and Contents.* Ed. Nancy Ide and Jean Véronis. 1995. 211-231.

Huitfeldt, Claus, and C. M. Sperberg-McQueen. "Representation and Processing of Goddag Structures: Implementation Strategies and Progress Report." *Proceedings of Extreme Markup Languages 2006.* 2006. <http://www.idealliance.org/papers/extreme/proceedings/>

Sperberg-McQueen, C. M., and Claus Huitfeldt. "Concurrent Document Hierarchies in MECS and SGML." *Literary & Linguistic Computing* 14.1 (1999): 29-42.

Sperberg-McQueen, C. M., and Claus Huitfeldt. "TexMECS: An Experimental Markup Meta-language for Complex Documents." 2001. <http://decentius.aksis.uib.no/mlcd/2003/Papers/texmecs.html>

Sperberg-McQueen, C. M., and Claus Huitfeldt. "GODDAG: A Data Structure for Overlapping Hierarchies." *DDEP-PODDP 2000.* Ed. P. King and E. V. Munson. Lecture Notes in Computer Science 2023. Berlin: Springer, 2004. 139-160.

W3C. *Document Object Model (DOM) Level 1 Specification.* The World Wide Web Consortium, 2000. <http://www.w3.org/TR/REC-DOM-Level-1>W3C Recommendation, September 2000

W3C. Ed. J. Boyer. *Canonical XML.* 2001. W3C Recommendation, March 2001

From Bibliography to Timeline: Flexible Infrastructure Bears Fruit

Ian R. Johnson *(johnson@acl.arts.usyd.edu.au)*
University of Sydney

Heurist (HeuristScholar.org) – introduced at last year's Digital Humanities conference – is an online system which integrates internet bookmarking with standard bibliographic functions, and provides strong social bookmarking and workgroup management capabilities. These include workgroup-centred tagging, rating and data sharing (private and public wiki spaces, threaded discussions and automated notifications) as well as publication of dynamically-generated subsets of the database.

Heurist is aimed at an academic audience. Signup requires authorisation by the system owner and, unlike generic systems such as Delicious or CiteULike, depends on social constraints to promote data quality (users are identifiable rather than anonymous). Heurist is available as a free service to academics (sensu lato) worldwide. We plan to release it as Open Source software in the course of 2007, once we have developed peer-to-peer communication between instances.

In developing the bibliographic functions of Heurist we did not wish to follow the 'easy way out' solution adopted by almost all existing bibliographic systems; that is, treating each bibliographic entry as an independent record in a 'flat file' database, using 'generic' fields with different meanings in different record types and redundantly repeating data such as book and publisher details in every chapter record.

First, we developed a method of recording variant records using metadata instances (a technique we pioneered for the Electronic Cultural Atlas Initiative metadata clearinghouse – www.ecai.org), allowing for a completely flexible data-defined and instantly extensible structure which handles missing data and repeating fields with ease. Fields and record types are soft-coded within the database and can therefore be added instantly through a web form; data entry forms are created on-the-fly from the soft-coded definitions.

Secondly, we construct complete bibliographic records by 'normalising' data into separate records reflecting discrete entities, thus reducing data redundancy and all its consequent problems. For example, a book chapter will be composed of four records - a chapter, a book, a book series and a publisher. These four records are linked together by inter-record relationships to form a complete bibliographic reference for the chapter. Correcting the book information will thus correct it for all chapters within the book.

There are a number of spin-off benefits in Heurist's data structure. In addition to all the normal field types – variable-length text, integer, decimal, date, logical and controlled lists – Heurist has three special field types which reflect the needs of the data structure and the particular interests of the research team:

- Geographic objects: Heurist stores point, point and direction, circle, rectangle, (poly)line and polygon objects. Geographic objects can be entered through a browser-based digitising function embedded in the data entry forms. We believe Heurist is the only spatially-enabled generic bibliographic database in existence;
- File objects: Heurist implements a generic method for uploading and managing files of various types – images, text, spreadsheets, sound, video etc. Multiple files can be attached to any record in the database and files can be assigned specific roles (eg. logo, thumbnail, introductory sound bite). Files can be embedded in output;
- Inter-record relationships: Heurist has rich functionality for relating a record in the database to any other record(s) in the database. To do this it defines a special type of field – a relationship field – which is a pointer to another record in the database. Relationship fields may be open or restricted to point to a specific type of database record. During data entry relationship fields pop up a search into the database, filtered to the appropriate record type. Relationships between records in the database are then just another type of soft-coded database record containing two relationship fields and a relationship type field. Some tricky code handles addition of new relationship types and directionality, so that a parent-child relationship, for instance, is reversed to child-parent if viewed from the other side. Relationship records also provide a datestamp (the range of time over which the relationship applies) and all the annotation capabilities of other Heurist records (including tagging, personal notes, wiki pages, threaded discussion, personal bookmarking and searchability).

We have been able to use the relationship functions of Heurist, combined with its flexible record typing and geographic objects, for some novel applications. These include mapping a network of research centres, projects, grants, methods, sites, researchers and publications (the 3DVISA network – heuristscholar.org/resource/45089) and recording a set of historical events and their relationships to create a browsable view of history (Silk Road TimeLine - heuristscholar.org/resource/45317).

The underlying data for the Silk Road timeline is rendered as a browsable list of interrelated events. When an event is selected, all related events are listed, grouped by relationship type – the web of relationships can be navigated by clicking on one of the listed events to select it as the 'root' of a new list of related events. The related events displayed in the list are simultaneously shown on a TimeMap web map (www.timemap.net) which is automatically updated whenever the list changes. Links on the map and on the list allow navigation between them and direct access to the backend database records. The model can be extended to visualising a different slice of history simply by copying the web page and seeding it with a different starting record, or by linking a starting record to the generic "Timeline Nodes" event which we have created in the database (heuristscholar.org/resource/48600). Our next step in this project is to develop a visual timeline which is generated on-the-fly from the current list of events and also linked to the TimeMap map and the event listing.

In this paper I will show how Heurist and TimeMap have been used to build this generic collaborative content creation model and visualiser for historical events, illustrated through the example of the Silk Road timeline project, and discuss the underlying data structures. I will discuss planned developments in the data model we use to describe historical events, and our research on improved visual methods of entering and viewing the web of relationships between historical events (or other types of record), including timelie visualisations. I will also discuss specific applications of the methodology adapted to museums, visitor centres, online encyclopaedias and the classroom.

Heurist and TimeMap have been developed by the Archaeological Computing Laboratory at the University of Sydney, under my direction, as part of the Sydney Humanities and Social Sciences e-Research Initiative (Heurist) and with funding from the Australian Research Council, ECAI, MacquarieNet and other user groups (TimeMap).

Relationship Mapping for Art Education and Research

Unmil Karadkar *(unmil@csdl.tamu.edu)*
Texas A&M University
Neal Audenaert *(neal@csdl.tamu.edu)*
Texas A&M University
Adam Mikeal *(adam@tamu.edu)*
Texas A&M University
Scott Phillips *(scott@cs.tamu.edu)*
Texas A&M University
Alexey Maslov *(alexy@cs.tamu.edu)*
Texas A&M University
Enrique Mallen *(e-mallen@tamu.edu)*
Texas A&M University
Richard Furuta *(furuta@cs.tamu..edu)*
Texas A&M University
Marlo Nordt *(mnordt@tamu.edu)*
Texas A&M University

Introduction

The *catalogue raisonné*, or reasoned catalogue, has long been a standard tool for representing large art collections. A typical *catalogue raisonné* includes images of artworks along with descriptive metadata, commentary, and background information (often a biography of the artist) about the collection. More recently, technological and infrastructural advances (in particular, cheaper secondary memory, increased network bandwidth, computational power, and digitization technology) have enabled development of the digital *catalogue raisonné*. Some of these catalogues have been developed primarily to support the construction of print-based catalogues (Lanzelotte, et al., 1993; Gladney, et al. 1998), while others are intended for online use (for example, The Vincent van Gogh Gallery and Gemini G.E.L.). The *Picasso Project* (Mallen, 2006) has developed a digital catalogue raisonné containing 11,000 of Picasso's artworks, along with 7,000 biographical entries. It is the most complete and up-to-date collection of Picasso's prodigious body of work.

Building on the premise that the logical structures of the book do not support scholarly inquiry adequately (McGann, 1997), we are using the Picasso digital catalogue to facilitate scholarly work with art collections. Researchers, students, and teachers in disciplines such as art history, painting, drawing, history of art, and art appreciation deal with art collections. They analyze and critique individual works and compare and contrast these with other works. They identify similarities between pieces of art and trace threads of influence between artworks, artists, styles, materials, themes, and social, geo-political, or personal events. These scholars interpret artworks, identify missing links, and communicate their findings. In the context of the Picasso collection, we support scholars in expressing and visualizing the complex, multifarious relationships between artworks via a Web-accessible software interface.

Approach

In a series of informal interviews with faculty members from art education, history, Hispanic studies, art history, and with local K-12 art teachers we found a diverse set of needs, interests, and approaches to working with artworks in both education and research settings. One key theme running through each of these areas is the need to discover and present relationships between artworks, although the specific relationships of interest varied by discipline. The art history scholar wishes to investigate relationships between artworks displayed together in an exhibition or to study works composed when an artist was with a particular lover. The historian wishes to view art in the context of significant historical events, for example, artworks created while Europe was anticipating World War I. K-12 teachers are interested in identifying artworks that provide good examples of specific drawing or painting techniques, such as the two-point perspective or the use of complementary color schemes.

In addition to the interviews, we also attended sessions of two college-level art history survey courses. We observed that instructors typically showed one or two examples of artworks from different artists or art movements, discussing each for a few minutes. In subsequent interviews, the instructors explained that lack of time constrains their ability to include additional works. Creating thematic sub-collections based on the relationships discussed in class could alleviate this problem, enabling students to study additional examples of materials covered in the classroom. These observations of classroom interaction and feedback from educators and researchers have informed our enhancements for supporting the representation and visualization of diverse relationships over the Picasso project's artwork catalogue.

Picasso's works cover a broad range of themes, topics, and materials, thus presenting a rich substrate of artworks for building a network of semantically diverse, meaningful relationships. In addition to the image collection, the Picasso project includes extensive metadata related to these works, such as its place and date of creation, medium, dimensions, current location, as well as exhibitions and books in which it has appeared. We leverage much of this metadata to express relationships based on ownership, materials, patronage, or chronology.

Interactive Relationship Visualizer

The Interactive Relationship Visualizer (IRV), an interactive, Web-based application, enables visualization of relationships that exist between artworks in the archive. The IRV interface displays image sub-collections connected by the relationship of the viewer's focus. In addition, it presents connections that exist within artworks in the sub-collection as well as those with others in the archive, enabling users to navigate the intricately interconnected hypertextual web defined by these relationships. While browsing, the the display changes to reflect the dominant relationship being displayed. In order to express a rich set of relationships, we are augmenting existing metadata to include type (such as still life or portrait), art movement (cubism, fauvism, surrealism), and content (woman, nude, vase, mirror).

The IRV distinguishes two broad categories of relationships, "inferred" and "specified." Inferred relationships are those which can be expressed in terms of the metadata elements provided for each artwork. Some inferred relationships can be expressed in terms of a single metadata value, such as "artworks created with oil on canvas." Others require mapping a range of metadata values onto a higher-level concept and require definitions involving multiple metadata fields. For example, identifying "paintings created in Paris around the time of World War II" is a two-step process. The system must map the timeframe of World War II to a portion of the traditional calendar and locate paintings created during this time. It then selects from this set, those that were created in Paris Finally, relationships such as "expensive paintings" involve subjective, theory-driven, and potentially variable definitions. A price that would be expensive in one art market might be comparatively inexpensive in another. Inferred relationships provide a powerful mechanism for exploring, discovering, and expressing relationships between artworks that leverages existing metadata.

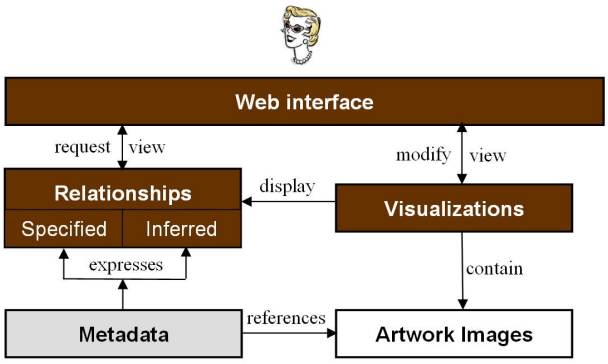


Figure 1: Interactive Relationship Visualizer system design

Figure 1: Interactive Relationship Visualizer system design.

In contrast to relationships inferred directly from existing metadata, other types of relationships must explicitly be stated. We refer to these as "specified" relationships. For example, Picasso sketched several rough drafts of large works, interspersed with smaller works. Thus, a chronological view of artwork around the time the *Guernica* was painted results in a sub-collection that includes these preparatory works as well as other, unrelated works. Hence, this "preparatory work" relationship must be expressed between early sketches of the *Guernica* and the final masterpiece. Another specified relationship is images based on a shared subject, for example, Picasso's interpretive series of works of Diego Velásquez's *Las Meninas*. Specified relationships afford us the ability to define and represent relationships between artworks that are difficult to derive from the descriptive metadata associated with each work. This category of relationships is critical for the expression of concepts based in established and novel analytical approaches to Picasso's work—allowing relationships based on information beyond that which is encoded in the collection. The drawback is that participation in specified relationships must be manually encoded.

Figure 1 displays the IRV system design. The Web interface employs specific visualizations for displaying different kinds of relationships. For example, the display of *Guernica* and its preparatory images uses a visualization that illustrates the centrality of *Guernica* relative to the other images displayed. In contrast, the display of all artworks in the *Las Meninas* series uses a table-like view, since no image is clearly central to this sub-collection. We employ artwork images from the Picasso project, reinterpret and extend existing metadata to express myriad connections between these artworks, and facilitate visualization of these relationships to support art scholars from various disciplines.

Future Work

We continue to add new metadata to enrich the relationships expressed in our archive. While new attributes enable us to express additional relationships, the growing number of relationships gets increasingly difficult to represent visually. We are investigating mechanisms to display secondary visualizations without overwhelming the presentation of the primary relationship views. As scholars analyze Picasso's works and life, the relationships of their interest are likely to increase in complexity as well as variety. It is not possible to express all imaginable relationships among these artworks a priori, nor is it possible to have all the necessary metadata. Enabling scholars to define useful metadata as well as supporting them in forming new relationships will engage them as partners in this project rather than as mere users.

The IRV has potential for exploration of artwork relationships in the classroom as well as for evaluating student performance via homework and papers. For example, students could explore a relationship and write a short paper about the artworks it encompasses. An instructor could create a relationship and ask students to identify the relationship embodied by the included artworks. Educators need assistance in the form of targeted features for successful use of the IRV in the classroom setting. We continue our dialog with instructors to channel the IRV's expressive power for enriching education.

Bibliography

Gladney, H. M., et al. "Digital Access to Antiquities." *Communications of the ACM* 41.4 (1998): 49-57.

Lanzelotte, R. S. G. "The PROTINARI Project: Science and Art Team up Together to Help Cultural Projects." *Proceedings of 2nd International Conference on Hypermedia and Interactivity in Museums, Cambridge, UK (1993).* .

Mallen, E., ed. *The Picasso Project* . Texas A&M University . Accessed 2006-10-06. <http://picasso.tamu.edu/>

McGann, J. "The Rational of Hypertext." *Electronic Text: Investigations in Method and Theory*. Ed. K. Sutherland. New York: Oxford UP, 1997. 19-46.

The Vincent van Gogh Gallery . . <http://vggallery.com/>

Washington DC: National Gallery of Art . *Gemini G.E.L.: Online Catalogue Raisonné*. Texas A&M University. Accessed 2006-09-07. <http://www.nga.gov/gemini/>

Done: "Finished" Projects in the Digital Humanities

Matthew Kirschenbaum *(mgk@umd.edu)*

Maryland Institute for Technology in the Humanities (MITH) and Department of English University of Maryland

William A. Kretzschmar, Jr. *(kretzsch@uga.edu)*

University of Georgia

David Sewell *(dsewell@virginia.edu)*

University of Virginia Press

Susan Brown *(sbrown@uoguelph.ca)*

School of English and Theatre Studies University of Guelph

Patricia Clements *(patricia.clements@ualberta.ca)*

Department of English and Film Studies University of Alberta

Isobel Grundy *(isobel.grundy@ualberta.ca.)*

Department of English and Film Studies University of Alberta

Done: "Finished" Projects in the Digital Humanities

Matthew Kirschenbaum

As the digital humanities continue to mature—theoretically, institutionally, as a set of critical practices—there will be an increasing desire to measure milestones, achievements, completion, and closure. While one instinct might be to bristle at such tendencies, and the traditional print-based scholarship they seem to imply or infer, being pushed to formulate a more fully realized set of responses can be a healthy exercise for a rapidly expanding field. "Done," as a shorthand convention that has quickly become ubiquitous in the culture of professional knowledge work, is emblematic in this regard, on the one hand unequivocal and absolute, on the other hand useful only in the context of low-level tasks that accumulate, often without finite bounds or comprehensive structure, in the service of some larger, less-defined endeavor. A server-side PHP upgrade may be "done," but what about the thematic research collection it exists to support?

So how do we decide when we're done? What does it mean to finish something? How does the "open ended nature of the medium" (a phrase we all pay lip service to) jibe with the reality of funding, deadlines, and deliverables? What can we learn from finished projects, both successful and unsuccessful? For that matter, how do we define success and failure? Are "we" the ones who ought to be defining it? If not, who? This panel attempts to begin formulating responses to such questions, by bringing together practitioners and project leaders from multiple institutional settings and contexts. William A. Kretzschmar's paper traces the local history of one particular project, grafting milestones and markers of what its gotten "done" to various versions and releases, a common enough convention in software and information and technology. His remarks raise issues not only for how one assesses closure, but also digital preservation and version control. David Sewell addresses the questions I've posed from the perspective of a publisher. His work at the University of Virginia Press's Rotunda digital imprint, itself unique in the field, offers a very different context, one where finding some measure of "done" is a financial and professional necessity. Finally, the team from the Orlando project offers the perspectives of one of the largest and longest running undertakings in the digital humanities, at the moment when the project has gone to press. The co-authors discuss not only what is done, but also what will stay done (and relevant) only by virtue of the ongoing expenditure of effort and resources, in the form of continual upgrades.

Large-Scale Humanities Computing Projects: Snakes Chasing Tails, or Every End is a New Beginning?

William A. Kretzschmar, Jr.

One of the motivating questions for this session is "What does it mean to finish something?" As it happens, the word "finish" can mean two things that have quite different implications for large-scale humanities computing projects. On one hand, according to the OED (sv) "finish" can mean 'To bring to completion; to make or perform completely; to complete.' On the other hand, the word can also mean 'To perfect finally or in detail; to put the final and completing touches to (a thing).' In my own work of this kind, the American Linguistic Atlas project (<http://www.lap.uga.edu>), we do neither of these things. We cannot come to an end of the work because we are witnesses and archivists of how Americans talk, and they keep talking differently across time and space. Neither do we think that our humanities-computing representation of our research is capable of being finally perfected, of achieving some perfect state, because the demands placed upon our research keep changing. If we view the entirety of the Linguistic Atlas Project as a "large-scale humanities computing project," the word "finish" is just not part of the deal. However, it is quite

reasonable to ask, as our granting agencies must ask, “what do you want money for this time?” or “did you accomplish what we gave you money to do?” From this viewpoint, the Atlas Project consists of a series of particular tasks or experiments, each one of which is capable of being “finished” in both senses of the word. In this paper, I would like to discuss the reality of funding, deadlines, and deliverables, as they relate to the sequence of tasks that make up the larger Atlas Project. In so doing, I hope to show the special character of work done deliberately as part of a sequence for a large-scale project, as opposed to work proposed as a singular task. The contextualization of the separate tasks leads to special cases of what it means to “finish” the work in either sense. The point of what follows is not the Atlas Project itself, but instead the way that individual tasks respond to the technical and academic situation at the time, and how our work and thinking over the years must change so that we can avoid the charge of being the snake that chases its own tail.

The Linguistic Atlas Project Web site has been notable over many years for its twin goals of interactivity for research (including the use of GIS) and making its data sets accessible and available to the public. I first programmed a GIS system on the Mac platform for our Linguistic Atlas data in 1990 (Kirk and Kretzschmar 1991, 1992; Kretzschmar 1992). It was widely used for teaching and research on American English in the early 1990s, and it immediately led to breakthroughs in how we were able to think about language variation data (Kretzschmar 1994, 1996; Kretzschmar and Lee 1992). The immediate task was “finished” in both senses, but the larger Atlas Project required more developments. While the Mac system was widely used, it was limited by restricted storage available to distribute data sets (at that time, chiefly through diskettes).

As the next task in the sequence, we then ported the system to the Web, which I first demonstrated at a conference in 1996 (Kretzschmar 1996b). We had been working on an interactive ftp/gopher system as early as 1994, but when Web technology became available we saw that it enabled perfectly what we had been attempting from another direction. The Web allowed us to make all of our textual data available, with many additional GIS features for locating speakers and information. The Atlas Web was a significant advance for both teaching and research (Kretzschmar 1997, 2002a), in line with the goals for an electronic atlas first set forth nearly a decade earlier (Kretzschmar 1988). After a while, we wanted to do new things, so we began work on a major revision of the Web site that finally came on line in 2003, which added even more interactive choices such as more flexible searches and tallies of the speakers and language data set. However, we kept "The Old Site" as a link on the new one, so that long-time users would find what was familiar to them, and also for users who did not want the greater complexity that came with greater flexibility of use. We could not just move it, however, because “The Old Site” had to run on a new platform and had to be compatible with the extensive Python scripting that ran functions on the New Site. The task of importing the Mac-based GIS system to the Web was complete by 1996, but not finished until 2003 with the platform change and the final touches of the more flexible site.

Still, the larger Atlas Project is nowhere near at an end. We are now rethinking what the site should do, from a text-based system to one that features audio and stored images along with text and GIS. This change has become possible only in the last two years, as much greater network-attached storage has become available (measured in Tb, before long Pb). We are now the one of the largest clients at the U of Georgia institutional storage array (which we share with bio-informaticists, physicists, and others usually considered to be power users) because of our archival audio files, and we are just at the beginning for audio and images. We now conceive of our new interviews as conversational corpora, in which text transcriptions serve as time-linked indices to audio files (Kretzschmar, Barry, and Kong 2005; Kretzschmar, Anderson, Beal, Corrigan, Opas-Hanninen, and Plichta 2006). While many users will want to listen to our speakers, others will want to perform acoustical analyses, now a strong trend in language variation research, as we ourselves now perform them (Kretzschmar, Lanehart, Barry, Osiapem, and Kim 2004; Kretzschmar, Kim, and Kong 2005). Our next task is to integrate sound with text and to enable acoustical research functions, while maintaining our interactive GIS functions—a whole new set of tools and problems from the previous task (Kretzschmar 2002b). New options for both hardware and for text encoding make it possible to consider the new site, just as the Web was a new option for the previous task.

So, are we "finished"? Yes, with the GIS Mark 2 of "The Old Site," but never finished either with keeping that site and its successors available or with new approaches to the information we keep as new technical possibilities and research demands appear. We can complete particular tasks, and often we can even “finish” particular tasks in the sense of polishing them for improved use. Yet one experiment does not make the whole line of research. One research proposal does not make the whole research program. While we can succeed with each task, we must always see tasks as part of the larger process that must continue if future users are still going to be able to take advantage of our resources. We would be the snake chasing its tail if all we did was keep polishing a single task, but we do well to make every end a new beginning as new possibilities become available.

The implications of the Atlas Project situation for other projects are definite and important. First of all, each project needs to consider whether it wants to be a single task, or whether it will be one of those continuing efforts for which any particular site

or tool is just a single stage in a larger, longer process. The reality of funding, deadlines, and deliverables makes it easier to propose and defend the single tasks, but in many cases we will be more honest, and better off in the long term, if we recognize that what we are doing is evolutionary rather than singular. We may have to work harder to convince funding agencies to support long-term work like the Atlas Project, but we must do that nonetheless. At the same time, our experience shows that we can indeed identify separate stages in our work that have independent value, and are thus fundable under the demands of deadlines and deliverables. Indeed, the demands of funding agencies can help those of us with long-term projects to organize our overall effort so that we keep it vital, interesting, and in the forefront of what new technology and standards can offer.

Bibliography

Kirk, John, and William A. Kretzschmar, Jr. ""The Analysis and Interpretation of Dialect Databases by Interactive Mapping"." *ACH/ALLC Conference, Tempe, 1991*. 1991.

Kirk, John, and William A. Kretzschmar, Jr. "Interactive Linguistic Mapping of Dialect Features." *Literary & Linguistic Computing* 7.3 (1992): 168-75.

Kretzschmar, William A. Jr. ". Linguistic Theory and Computer Modeling of Linguistic Survey Data." *ACH/ALLC, Paris, 1994*.

Kretzschmar, William A. Jr. "Quantitative Areal Analysis of Dialect Features." *Language Variation and Change* 8.1 (1996a): 13-39.

Kretzschmar, William A. Jr. "The LAMSAS Internet Site." *NWAVE, Las Vegas.* 1996b.

Kretzschmar, William A. Jr. "Teaching American English Online." *Journal of English Linguistics* 30.4 (2002a): 318-327.

Kretzschmar, William A. Jr. "TEI and Linguistic Interviews." 2002b. <http://www.tei c.org/>

Kretzschmar, William A. Jr., Jean Anderson, Joan Beal, Karen Corrigan, Lisa-Lena Opas-Hänninen, and Bartek Plichta. "Collaboration on Corpora for Regional and Social Analysis." *Journal of English Linguistics* 34.3 (2006): 172-205.

Kretzschmar, William A. Jr., Betsy Barry, and Nicole Kong. "Publication of Full Interviews from the Atlanta Survey Project." *ADS/LSA 2005, Oakland.* 2005.

Kretzschmar, William A. Jr. "Interactive Computer Mapping for the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS)." *Old English and New: Essays in Language and Linguistics in Honor of Frederic G. Cassidy*. Ed. N. Doane, J. Hall and R. Ringler. New York: Garland, 1992. 400-14.

Kretzschmar, William A. Jr. "Computer-Assisted Study of American English Lexical Data." *In From AElfric to the New York Times: Studies in English Corpus Linguistics.* Ed. Udo Fries, Viviane Müller and Peter Schneider. Amsterdam: Rodopi, 1997. 239-47.

Kretzschmar, William A. Jr., MiRan Kim, and Nicole Kong. "Vowel Formant Characteristics from the Atlanta Survey Project." *ADS/LSA 2005, Oakland.* 2005.

Kretzschmar, William A. Jr., Sonja Lanehart, Betsy Barry, Iyabo Osiapem, and MiRan Kim. "Atlanta in Black and White: A New Random Sample of Urban Speech." *NWAVE 2004, Ann Arbor*. 2004.

Kretzschmar, William A. Jr. "Computers and the American Linguistic Atlas." *Methods in Dialectology: Proceedings of the Sixth International Conference on Methods in Dialectology*. Ed. Alan Thomas. Clevedon: Multilingual Matters, 1988. 200-24.

It's For Sale, So It Must Be Finished: Digital Projects in the Scholarly Publishing World

David Sewell

Since late 2004, the University of Virginia Press has been offering as part of its catalog a group of scholarly publications that exist in online format only. Distributed under our Rotunda imprint (<http://rotunda.upress.virginia.edu>), these publications are a mix of born-digital, digitized print, and hybrid creations in digital humanities and social sciences. Some of them began independently as self-published projects, usually under the auspices of one or another digital center such as the Maryland Institute for Technology in the Humanities or the Electronic Text Center at the University of Virginia; others were initiated by the Press. Whatever their origins, once accepted as Rotunda publications they have all been subject to conditions of production similar to those for our printed books, namely contracts, peer review, marketing campaigns, and agreed-upon scope and deadlines. Clearly the academic marketplace offers an extrinsic definition of a finished digital project: if it's for sale it must be done.

From this point of view, there is nothing inherently different between print and digital scholarship. Academic publishers have long been in the business of imposing admittedly arbitrary conventions, limitations, and deadlines on masses of scholarly discourse in the interest of presenting them as discrete units that can be offered in the marketplace as completed products (not just for purchase, but also for authoritative review and citation). And this always entails negotiation between the author's ideal creation (which may well be both theoretically unbounded and technically unachievable) and the publisher's available time and resources. The result is a pragmatic compromise, one that has worked reasonably well for both

authors and readers. I would argue that the yoking of digital humanities projects to theories of the open-ended text is in many ways an unfortunate by-product of their initial emergence during the ascendancy of postmodernist theory, with a glorification of the unfinished that may harm their credibility. When a publisher issues a scholarly article, book, or digital publication, its status as "finished" represents a social contract that the necessary stages of peer review, editing and design, and quality assurance (proofreading, plus user testing for digital work) have been performed.

Here is one possible typology of digital projects according to their degree of completion at the time of publication:

1. Self-contained monograph-like objects. All content and functionality that is ever intended to be part of the project is present at first publication. Future updates are expected only for corrections and migration as required to new software environments
2. "Version 1.0" digital projects. All substantive content is complete at first publication, but not all planned aspects of the user interface. Future updates are expected to add functionality that was not originally available.
3. Series-like projects. Content is added in discrete stages; it may be quite large in extent, but is definable. All intended functionality may be present in the first installment, or may expand as in category #2. Print analogues would include literary or documentary editions published in multiple volumes over a number of years or decades, and ongoing reference works where new and updated entries are added periodically.
4. Truly open-ended projects. The nature of the content, subject matter, and/or authorship is such that no particular state of the project could ever be meaningfully said to represent "completion."

To date, Rotunda has published digital projects that fit into each of the first three categories. Their work flows and timetables have been more fluid and sometimes more tentative than for the Press's printed books, but not different in their basic nature. It has always been possible to define what will constitute a finished project, outline the steps necessary to arrive at it, and establish reasonably solid deadlines for the process.

I will provide concrete examples for several Rotunda publications of decisions we have made in conjunction with authors about what could or could not be accomplished by a deadline, the distinctions between Version 1.0 and Version 2.0 features, and the kinds of updates we have found ourselves needing to make to finished publications. I will also discuss why the truly open-ended projects of category 4 are more problematic for traditional scholarly publishers; our experience suggests that they are the most likely to require new mechanisms and/or institutions for quality control and sustainability.

Orlando Done! The Tension Between Projection and Completion in Digital Scholarly Research

Susan Brown

Patricia Clements

Isobel Grundy

Orlando: Women Writers in the British Isles from the Beginnings to the Present was published on the world wide web, by subscription, by Cambridge University Press in June 2006 (Brown, Clements and Grundy). Its freshly researched, scholarly, literary-historical prose is encoded in an XML application designed for this purpose by the Orlando research team and repeatedly revised and tweaked during the course of research and writing. The searching functionality and the production architecture were created at a later stage but largely by the same team.

The project's cofounders were new to digital humanities research, so our notions of scholarly process and completion related to conventional print publications. As Claire Warwick has noted, the idea of what is "complete" or "publication-ready" in academic culture has emerged from a complex set of human factors relating to such matters as the attribution of credit by institutions and funding structures, as well as the conception of what is required intellectually for a product to be done (Warwick 368). Such factors undoubtedly entered into how *Orlando* got done. Although our projected milestones for the project broke out a number of steps, our research plan and our proposal to our funding agency had projected a single moment of completion when the planned electronic history would be ready alongside several related print volumes of scholarship. Once underway, we found getting done more challenging than anticipated, as the enormity of what we had undertaken on the technical side became evident. This is clearly a risk of methodologically experimental research of any kind, and particularly relevant to digital humanities work. One thing that seems crucial in the design of digital humanities projects is to design projects modularly, with a number of discrete and in some way publishable deliverables. *Orlando* struggled for funding in later stages as a result, we believe, of research design that end-loaded deliverables, combined with a funding environment that was not attuned to digital humanities work.

It soon became clear that it would be necessary to stage the completion of the project. We uncoupled the electronic from the print publication in terms of timing, so that the former stands alone initially, even though we will integrate the print volume material, now being written, with the existing published textbase. Although over the course of *Orlando*'s development we felt some pressure from both prospective users and from

the digital humanities community to publish at an interim stage a small portion of the electronic materials, we held back for two major reasons. First, we had a strong sense that there were intellectual demands for a critical mass of materials with a certain degree of coverage. We wouldn't be "done", for instance, without having completed the materials on Virginia Woolf and, because much feminist work has resisted the establishment of a small canon of female writers at the expense of others, Woolf needed to be situated in relation to her less prominent contemporaries. Staging releases of material is increasingly common for digital electronic projects, but we would contend that a certain critical mass remains necessary to establish scholarly confidence in quality of a resource, although what that critical mass is will differ from project to project. Secondly, because our customized tagset required us to build a fairly complex XML delivery system, we felt we needed a relatively "complete" interface, with some usability testing, that could demonstrate to users some strengths of the markup into which the project invested so much time and intellectual labour. For these strengths of the system to be apparent, we again needed a critical mass of materials for hyperlinking and search results. Research conceptualization and publication options are thus both crucial determinants of what "done" will mean for a particular project.

Orlando's decision to publish with an online press, in one sense made it easy to know when the textbase was done. Published is traditionally done (though even then, glitches in getting the textbase up and running on the publisher's server meant that we actually celebrated the publication some days before it happened). But published electronic projects don't get put on a shelf in a library. Being unconstrained by the materiality of print reinforces the arbitrariness of deciding that something is done in the sense of complete, defined in the *Oxford English Dictionary* as "Having all its parts or members; comprising the full number or amount; embracing all the requisite items, details, topics, etc.; entire, full." In this sense, *Orlando* remains undone, and our contract with our publisher recognises this fact by stipulating for updates.

Regular updating of content (more entries, expansion of existing entries, more contextual material) has already, therefore, begun with an update adding new material in January 2007. This is an obvious publicity tool. Enhancement of functionality (quite apart from action rendered necessary by, for instance, the advent of a new version of Internet Explorer), though it may be less productive of instant, visible returns than enhancement of content for most users, is nevertheless going to be vital for achieving the status of a first-stop general reference work, and for retaining its currency in the evolving electronic environment.

This continuing updating of both content and functionality means that the core team must be held together and the project must continue to find funding. As part of its strategy of sustainability, and with the purpose of giving *Orlando* a continuing institutional identity, the textbase has been licensed to the University of Alberta. This gives a much broader constituency an interest in the project's success, and ensures that the team of this moment forms a continuity with the future as well as the past. It would be easy to cast all of this proliferating responsibility somewhat negatively, as though continuing work on an already-published project were a burden, but the fact is that all of the previous research, writing and encoding that has been done towards publication will work together with user feedback to enable new and newly informed research. Both expansion of content (even into new areas) and enhancement of functionality represent continuing research potential, even while they ensure that "done" can – need? -- never be anything but a figure of speech.

Apart from the issues of sustainability, the degree of *Orlando*'s long-term success still hangs in the balance. Publication is, potentially, the beginning of the road to success, not its end. We would define success not merely as the selling of subscriptions (though obviously *Orlando* needs to do well in the marketplace if a substantial portion of its funding needs are to be met from royalties), but also as the establishing and maintaining of a pattern of heavy use of the textbase by its subscribers: indeed, a pattern of reliance on it as the first stop for either specialist or general information. We would further define it as having users exploit the broad range of possibilities represented by the encoding.

This abstract, written four months after publication, is in every way provisional, but it is clear from early (and so far highly encouraging) user feedback in the form of emails and a single brief review that for most of our users the electronic side of the textbase is ancillary to the content. They comment on content, apparently assuming that the new resource should stand or fall on content alone. Nevertheless, those who built the textbase understand that this feedback must be to a large extent influenced by the users' responses, not directly articulated and perhaps not even consciously formulated, to *Orlando*'s functionality.

At this early stage most user messages come from the literary-specialist part of an audience which will hopefully also take in computing specialists and general readers. Most messages comment on the coverage of individual authors (or very particular groups of authors), on whom the commenting user tends to be an expert, so that feedback is skewed both towards the literary and towards the individual author entries, which are in electronic terms the least sophisticated part of the text. Several user messages have praised the "links". This likely means the hyperlinking of words tagged as names, organization names, titles, or places. Such linking is an obvious and elementary feature of myriad electronic texts. It seems therefore that users who said they appreciated the links probably had in

mind the way that the encoding sorts and organises the contexts in which linked words are used, and the very particular design of the links screen to convey this information. Messages which praise the searching, or the navigation, are even harder to decode.

We look forward to more comment on functionality and production, and further reports from general-interest users who have begun, for instance, by looking at coverage of gardening, as well as to a breakdown of statistics about kinds of use. In addition, we want to undertake standard usability testing. Log analysis would also be invaluable if we could overcome some technical hurdles. This sense in which work on *Orlando* remains "undone" should be common to most digital humanities projects: if we are to evolve useful tools and resources, both our own and those to come, we need carefully to assess how people approach and use them.

This project directs its research towards two practically inexhaustible fields (women's literary history and the capacity of computing, and specifically of extensive XML markup, to serve the needs of this humanities topic). Neither of those fields becomes closed to further investigation by the fact that the project is "done" in the sense that it has reached the public. For this project, so closely focused on its major deliverable, the new, post-publication phase has simply opened a third area of enquiry: that of the relations between *Orlando* and its users.

Bibliography

Brown, Susan, Patricia Clements, and Isobel Grundy. *Orlando: Women's Writing in the British Isles from the Beginnings to the Present.* Cambridge: Cambridge University Press Online, 2006. <http://orlando.cambridge.org>

Warwick, Claire. "Print Scholarship and Digital Resources." *A Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing Ltd, 2004. 366-82.

Semantic Clustering in the Wild

Aaron Krowne *(akrowne@emory.edu)*
Emory Univeristy
Alice Hickcox *(alice.hickcox@emory.edu)*
Emory University
Stephan Ingram *(frowey@gmail.com)*
Emory University

Problem Space and Context

The Southern Changes Digital Archive is a digital collection of 25 years of Southern Changes, a publication of the Southern Regional Council.[1] The Beck Center of Woodruff Library has digitized the issues of Southern Changes and made them available as a fully searchable text archive. [2] The collection consists of 25 volumes, over 100 issues, containing nearly 1000 articles.

Emory University is participating in the Digital Library Federation's Aquifer Project and the Southern Changes Digital Archive is one of the bodies of work that we proposed to include. As we inspected the elements of the MODS Aquifer schema,[3] it became clear to us that every item in the metadata collection would end up with identical labeling, since the collection lacked subject assignments for the individual articles.

Thus, the challenge that we faced, in order to make this material useful to users of the MODS-Aquifer metadata, was to assign subject classifications to each of the articles in the collection with some level of automation and a minimum level of human intervention (given the ~1000 individual articles). We also realized that this undertaking would help to make the collection more browseable by providing a subject-based access point into the digital library. To achieve these ends, we turned to the tools and techniques developed on the MetaCombine project here at Emory.

Semantic Clustering in the DL Production Environment

Semantic clustering is the notion of using computing techniques to group formerly unorganized artifacts based on latent aspects of meaning. For library information, this translates into grouping texts and metadata records by their

inherent “topics.” Semantic clustering techniques have come far in recent years, and hold great promise for enriching many information applications. However, there is a wide gulf between the state-of-the-art research in this area, and its uptake in real-world applications.

A major reason for this disjunct is that research systems for semantic clustering (and related techniques such as classification) are not fit for the production environment: they are too ad hoc, too brittle, too encumbered by intellectual property constraints, too specialized, poorly organized, and they typically require programming to deploy in the local computing environment and adapt into situational use.

But there is another factor that is more interesting for this study: that the automated aspect of semantic clustering covers only part of the production process; as we found on the MetaCombine [4] project (and which many others confirm), some human refinement is always needed. However, tools and facilities which mesh the automatic with the manual to perform this refinement seem to be unavailable.

In sum, there exist few (if any) turnkey systems that are accessible to digital librarians which allow the integration of semantic clustering into the DL environment.

We began taking steps to address these problems on the MetaCombine project, by adding to core semantic clustering systems a some HTML-based report, a visual browsing application and a “scheme editing” system. All of these systems operated on the "raw" outputs of semantic clustering, and served to help comprehend, navigate, and (in the case of the latter) manually refine the output.

The central specific problems we found with semantic clustering results in MetaCombine [5] were:

- redundant clusters
- confusing and otherwise "rough" cluster descriptors (labels)
- spurious clusters
- questionable classifications

The scheme editor was created to allow the revision of a cluster hierarchy to ameliorate the above issues, helping to make it a more true category (or subject) hierarchy, fit for end-use applications (such as subject tagging and subject-based browsing).

We have now begun the process of using these tools to work semantic clustering techniques into production digital library services. The focus of the present paper is one of these real-world test cases: the Southern Changes collection and digital library. However, key for our investigation is that the library production environment itself is a major point of interest; thus this work represents more of an exploration of the application of the MetaCombine tools to this environment, rather than another quantitative test of clustering or classification accuracy for the particular algorithmic kernel.

Our Study

We have approached the Southern Changes clustering investigation as an informal study. This is both because the needs of the Southern Changes project are practical and because this investigation does not have the scale to have the form of a controlled experiment or series of experiments.

The structure of the study is:

1. interface the collection with the clustering system
2. perform clustering with a variety of parameter settings
3. explore and refine output in MetaCombine tools (HTML reports, scheme editor, visual viewer).
4. implement refinements to clustering (tweak parameters, fix bugs)
5. implement refinements to MetaCombine tools programs
6. go to step 2.

The study thus has an iterative nature and is open-ended, and as we go through these iterations, we are learning much about the nature of clustering and the contextualization of the task in the digital library environment. As indicated above, our clustering tools are also evolve throughout, moving beyond research to practice.

The insights gleaned from this process form the core of our reporting for this forum. The types of items we intend to report on, along with initial impressions include (but are not limited to):

- The utility of the scheme editor, HTML reports, and other tools

Thus far, these tools have proven quite useful for determining the latent topics discovered by the clustering engine, and give a basic level of understanding of what kind of documents are assigned to each. However, it remains difficult to get a more complete or holistic sense of how well a large number of documents has been assigned.

- The difficulty of deploying the tools

Thus far, the difficulty is moderate, for a librarian with some programming experience. Some problems in deployment of the scheme editor tool were isolated and future distributions will have fixed them; so the difficulty level is improving.

We have not broached the topic of re-deploying the clustering kernel itself for this project. However, this may not ultimately be necessary, as we are pursuing a federated model for this and other metadata enhancement services.

- The utility of clustering itself

Flat (one-level) clustering seems to do quite well with the right parameters and with the documents for which there is the most confidence. Fidelity seems to deteriorate rapidly for hierarchical clustering past the first level (with the first of two hierarchical clustering methods), though it is still unclear if this is more a result of the collection or the clustering system. Our second method of hierarchical clustering needs further development to be fully evaluated in end-use.

We believe we have established that the multiclassification functionality of the system is extremely useful; as topical ambiguity is compensated for by the multiple labels. This makes intuitive sense, as there typically is no single categorization for an interdisciplinary humanities article.

It appears significant work needs to be done in thresholding classification assignments, as well as making the determination of if and when confidence is high enough to consider a record “classified”.

- The feasibility of integrating outputs of clustering into production systems

As described in more detail below, we believe there will be little difficulty in utilizing the outputs of this system, once we have determined that output is satisfactory.

Given the proliferation of XML handling tools, at minimum, our current, simple XML format for representing a clustered collection organization should be easily translatable into our existing digital library system. The most challenging part may in fact prove to be combining the generated subject tags with the existing TEI metadata structure.

Implementation Plans

The results of the clustering are XML files that contain the renamed category, associated with the article by means of the id attribute assigned to the article. We will apply an XSLT transform to add an attribute that indicates the subject(s) assigned to each article, using the TEI "ana" attribute. We can then harvest the records with the subject information, and as a further step, add a subject browsing capability to the web site, using the "interpGrp/interp" elements. This will enable dynamically generated categories for searching and browsing.

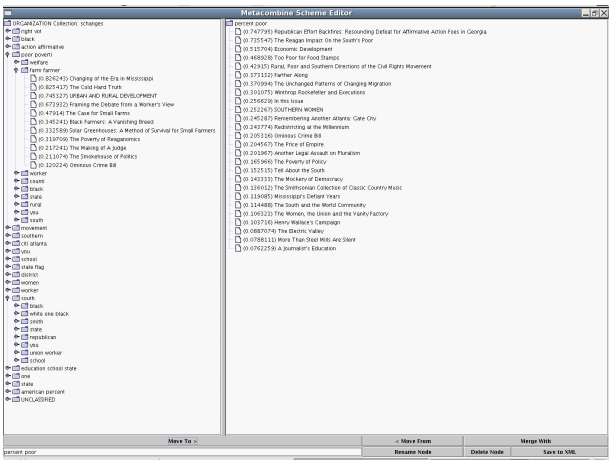


Figure 1:

1. Southern Regional Council <http://southerncouncil.org>
2. Southern Changes Digital Archive <http://beck.library.emory.edu/southernchanges/>
3. Aquifer MODS Guidelines <http://www.diglib.org/aquifer/dlfmodsimplementationguidelines_finalnov2006.pdf>
4. MetaCombine project site <http://www.metacombine.org/>
5. MetaCombine Final Report <http://www.metacombine.org/reports/project/MetaCombine-Final-Report.pdf>

Digital Representation and the Hyper Real

John Lavagnino *(John.Lavagnino@kcl.ac.uk)*
King's College London
Willard McCarty *(willard.mccarty@kcl.ac.uk)*
King's College London
Susan Schreibman *(sschreib@umd.edu)*
University of Maryland Libraries

Art has never been a mere mirror up to nature, yet as in no other medium has it been so easy to create a simulacra of reality as with digital technology: a 'heterocosm', both simulating the familiar while deconstructing it. This session brings together three theoretical papers that explore how mimesis might be used as a paradigm from which to explore the relationship between the digital world, the analogue world, and the space between them; digital surrogates and their analogue counterparts; how familiar terms like object, imitation, copy, original function in the digital realm; and the notion that a digital representation may be more appropriately termed a simulacral identity, reflecting, not the object itself, but our beliefs and conventions about it. This session will explore digital representation as conscious fashionings of hyper-reality, computational zones or subspaces which employ the unreal and non-existent to recreate the material world, pointing to the past, and the future, in unexpected, fresh, or subversive ways

Being Digital, or Analogue, or Neither

John Lavagnino

When I speak of something being digital or analogue, I draw on a connected pair of terms that is often thought to cover all possibilities in a way that's theoretically well grounded. But the origin of these two terms is not theoretical, and the conventional opposition reflects a pragmatic view of approaches to making computing machinery rather than deep and inherent qualities of information. The two terms also do not exhaust the world: most things are neither digital nor analogue, because those terms describe information that has been carefully prepared for machine processing. Nonmachines get by in the world without that restriction of input.

The separate ideas of digital and analogue representation come not from theory but from engineering. Digital representation has its background prior to the days of the digital computer, in the longer history of numerical calculating machines, and in particular in the use of punch cards and tabulating machinery, which could not only perform computations but provided ways to manage large bodies of information. Analogue representation goes back to a different tradition of calculating machinery, in which physical operations that could be interpreted as performing computations mattered much more than storage of information in any quantity. In the mid-twentieth century, a moment came when both approaches had significant applications, and two separate strands of technological development became a pair of options, both sometimes applicable to the same tasks (Wiener; Mindell). Today the two are commonly thought to exhaust all possibilities; but at the same time there is a marked status hierarchy, as digital systems continue to spread everywhere and analogue systems have a minor or nearly invisible existence in unexciting devices like thermostats. On this view, the digital has the prestige of being made by us, and the analogue has the consolation of covering everything that isn't manmade and a few things that are. Or, in one very common version of this opposition, thought is digital and reality is analogue.

Although many people assume in this way that everything is either digital or analogue, most things aren't actually either, because most things are not information prepared for machine processing. All our machinery for processing information, analogue or digital, has the common property of ignoring all but a restricted slice of the world as we experience it, and having no way to notice phenomena beyond that slice. We craft these machines to work with particular inputs and sometimes learn to change our behavior so that machines get the right sort of input. Representations made for other purposes (art, in particular) are based on selections, too, but because they are not created to serve as a basis for computation they don't present the same kind of claim to be authoritative versions of reality. Nothing is digital or analogue until we actually get it into a computational system; these properties result from our choices about how to process reality, and are not inherent in reality itself.

If reality is not by default analogue, thinking is not by its nature digital. Much work in artificial intelligence has been based on the hypothesis that the brain and mind work like digital computers: so that, while our current systems might be crude, they are still on the right path of development towards the real thing, and it's important that the real thing is digital. The hypothesis has been highly productive as a basis for useful work, but is at odds with the biological account of how the brain works; John von Neumann argued that the brain's machinery was essentially different from that of both digital and analogue computers. That difference is unsurprising, since machines of other kinds work very differently from biological systems with comparable functions. The natural form of technology “is typically tiny, wet, nonmetallic, non-wheeled,

and flexible; human technology is mainly the opposite: large, dry, metallic, wheeled, and stiff" (Vogel 271).

In one classical line of discussion of the digital-and-analogue pairing, Nelson Goodman's, written texts serve as an example of the digital mode: on this view, written symbols are intended as discrete and unambiguous tokens chosen from a fixed set (142). The way that texts lend themselves so naturally to transformation into electronic form may seem to provide evidence for this view: that they're not only readily made into usable digital objects but that they inherently *are* such objects. But in Goodman's account the digital nature of the alphabet is an idealization: you are supposed to be able to tell your letters apart unambiguously, yet we get by working with handwriting and bad print that fails in this regard. Much more significantly, we notice other things than the choice of token; as many accounts of communication point out (Roman Jakobson's, for example), there are many functions of verbal messages besides the transmission of the information. Digital representation in computers is different because they are designed to recognize nothing but the choice of token; what is an idealized account when applied to human use of the alphabet is a perfect account of computing, because it describes how computers are built to work.

Conventional accounts of digital representation rarely mention how much of the workings of a computer are there to keep the digital data digital, to prevent it from being degraded by noise; works on electronic engineering never omit the point. We miss a key feature of the digital and the analogue when we think of them as static properties that happen to everything without effort; they are instead deliberate creations. They have been highly successful creations, representations of reality that lend themselves to many uses; we ought to recognize them as our creations, and not mistake them for natural phenomena.

Bibliography

Goodman, Nelson. *Languages of Art: An Approach to a Theory of Symbols.* Indianapolis, IN: Bobbs-Merrill, 1968.

Jakobson, Roman. "Linguistics and Poetics." *Linguistics and Poetics.* Ed. Thomas A. Sebeok. Cambridge, MA: MIT Press, 1960. 350–377.

Mindell, David A. *Between Human and Machine: Feedback, Control, and Computing before Cybernetics.* Baltimore, MD: Johns Hopkins University Press, 2002.

Vogel, Steven. *Cat's Paws and Catapults: Mechanical Worlds of Nature and People.* New York: Norton, 1998.

Von Neumann, John. *The Computer and the Brain.* 2nd ed. . New Haven, CT: Yale University Press, 2000.

Wiener, Norbert. *Cybernetics; or, Control and Communication in the Animal and the Machine.* Cambridge, MA: MIT Press, 1948.

Looking Backward, Figuring Forward: Modelling, its Discontents and the Future

Willard McCarty

Alan Turing's scheme has not been adequate to computing "in the wild" for more than 50 years (Mahoney 1997: 621), but it does have two fundamental implications for work in the humanities. Its first implication is that intellectual gain from the computational analysis of a cultural artefact comes primarily from comparing it to its digital representation as this is improved through repeated trials and adjustments. Its second implication is that in principle there can be no limit other than human ingenuity to the forms computing can take. Hence computing's basic tradeoff: on the one hand, reduction of the artefact to computational form guarantees a permanent though changing gap between its transcendent reality and its calculable representation; on the other, the mutability of computing allows for no end to the perfective attempt to reach the former with the latter. This attempt I have called "modeling" (McCarty 2005).

The proposed paper takes the centrality of modeling for granted, but it attempts to move beyond the inherent limitations of a process that by definition only imitates. Modeling is directed to a pre-existing conception of an artefact one wishes to study; its strength is in contesting that conception in comparison against one's best attempt at representing it rigorously. The discrepancies it discovers may well be, in Jerome McGann's words, "the hem of a quantum garment" that trails into our future (2004: 201), but modeling gives us little help in imagining that future. Turing's scheme guarantees innumerable forms of computing, but how best are we to work toward them? Although, as Edsger Dijkstra remarks in his contribution to *Beyond Calculation: The Next Fifty Years of Computing*, it may seem "utterly preposterous" to predict this future, as teachers we do it all the time in deciding what to teach, what to ignore (1997: 59). As researchers we get hints of the future, or hopes for it, when no existing data model, or way of using computers, will do – when (to take an example from my own work) neither textual encoding nor relational database design satisfies, and we are left with a hunger for something other than what we have. Can we do better than such backward looking glances into the future? Can we imagine it directly?

One answer is supplied by Empirical Modelling (EM), presented to the last North American ACH/ALLC conference (Beynon, Russ and McCarty 2006) and further articulated in a recent MSc dissertation (King 2006). EM focuses on the present and presence of tacit experience, which is as close to the future as

we ever get. Another answer comes out of work in critical theory, e.g. by N. Katherine Hayles, whose focus on writing, and cultural productions generally, lifts the gaze to what is emerging – to "emergent" phenomena, as they are known (1999). Taking clues from both, I propose to explore and talk about a third answer arising from reflective work in the history, philosophy and anthropology of the natural sciences. In the philosophy of physics, for example, Ian Hacking has argued that rigorously imagined entities are made real when we learn to manipulate them (1983). In the history of technology, Peter Galison shows that the devices we invent tend to pull us forward into conformity with them (2007) – an argument quite close to one Northrop Frye made at the first joint ACH/ALLC conference, citing such human inventions as the wheel and the book (1990). In the history of chemistry, Mi Gyung Kim examines how 19th Century researchers worked to establish the reality of their substances, suggesting a surprisingly immediate interrelation of the imagined and the real (2000). In biological anthropology, Terrence Deacon argues beyond the uncomfortable limitations of a mechanical world-view and strict Darwinian evolution to a new conception of teleology, "to identify a real and substantial sense of the 'pull' of future possibilities in terms of 'pushes' from the past" (2006).

In the proposed paper, I summarize this work in the natural sciences and use it to construct a theory of emergence in humanities computing. I base my exposition on the underlying argument that use of computing, with its emphasis on "how we find out, not… what we find out" (Hacking 2002), brings us into productive relation with the experimental sciences without in any way compromising our orientation to the humanities (McCarty 2002, 2006, 2007). Sumarizing my earlier work, I suggest briefly at the beginning of the paper how computing has allowed us to create within the humanities a computational zone or subspace, within which practitioners may treat cultural artefacts *as if* they were only data, and so apply to them *something like* natural law. I argue that the conjectural, as-if status of what may be done within this zone gives us a defensible way of importing powerful scientific conceptions, such as Hacking's realization-by-manipulation or Deacon's biological teleology, and apply them to our artefacts of study, not in order to test what we think we know but to imagine and realize what we do not know.

In earlier work I have used this conjectural relation with the natural sciences to ground the practice of modelling in its scientific past (McCarty 2005). Here I use it as entry-point to speculations on how humanities computing might lead the disciplines which it most immediately serves into a fruitful relationship with the sciences and so to an end of the epistemic wars foreseen by Richard Rorty (2000).

Bibliography

Beynon, Meurig, Steve Russ , and Willard McCarty. "Human Computing – Modelling with Meaning." *Literary & Linguistic Computing* 21.2 (2006): 141-57.

Deacon, Terrence W. "Emergence: The Hole at the Wheel's Hub." *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion.* Ed. Philip Clayton and Paul Davies. Oxford: Oxford University Press, 2006.

Dijkstra, Edsger W. "The Tide, not the Waves." *Beyond Calculation: The Next Fifty Years of Computing.* Ed. Peter J. Denning and Robert M. Metcalfe. New York: Copernicus, 1997. 59-64.

Frye, Northrop. "Literary and Mechanical Models." *Research in Humanities Computing 1: Selected Papers from the 1989 ACH-ALLC Conference.* Ed. Ian Lancashire. Oxford: Clarendon Press, 1991. 3-12.

Galison, Peter. *Building, Crashing, Thinking.* Forthcoming.

Hacking, Ian. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science.* Cambridge: Cambridge University Press, 1983.

Hacking, Ian. *Historical Ontology.* Cambridge, MA: Harvard University Press, 2002.

Hales, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics.* Chicago: University of Chicago Press, 1999.

Kim, Mi Gyung. "Chemical Analysis and the Domains of Reality: Wilhelm Homberg's Essais De Chimie, 1702-1709." *Studies in the History and Philosophy of Science* 31.1 (2000): 37-69.

King, Karl George. "Uncovering Empirical Modelling". Unpublished MSc dissertation. Computer Science Department, Warwick University, 2006.

Mahoney, Michael S. "Computer Science: The Search for a Mathematical Theory." *Science in the Twentieth Century.* Ed. John Krige and Dominique Pestre. Amsterdam: Harwood Academic Publishers, 1997. 617-34.

McCarty, Willard. *Humanities Computing.* Basingstoke: Palgrave, 2005.

McCarty, Willard. "The Imaginations of Computing." Richard W. Lyman Award lecture, National Humanities Center, Research Triangle Park, North Carolina, 6 November. 2006.

McCarty, Willard. "Humanities Computing: Essential Problems, Experimental Practice." *Literary & Linguistic Computing* 17.1 (April 2002): 103-25.

McCarty, Willard. "Being Reborn: The Humanities, Computing and Styles of Scientific Reasoning." *Renaissance Studies and New Technologies: A Collection.* Ed. William R. Bowen and Raymond G. Siemens. Tempe, AZ: Medieval and Renaissance Texts and Studies, Forthcoming.

Rorty, Richard. "Being That Can Be Understood Is Language." *Gadamer's Repurcussions: Reconsidering Philosophical Hermeneutics.* Ed. Bruce Krajewski. Berkeley: University of California Press, 2004.

Beautiful Untrue Things: The Digital Dilemma

Susan Schreibman

In Oscar Wilde's dialogue *The Decay of Lying*, Vivian and Cyril discuss the interdependence of Nature and Imagination, with imagination, in that typically Wildean fashion, more faithfully representing the real than the material world. For much of the dialogue, Vivian reads an essay that he has authored to Cyril: his thesis is that there must be a return to 'lying in art' (5): a return to the roots of art in the purely imaginative abstract. This imaginative work, rooted in the 'unreal and non-existent' takes as its rough material life, 'recreates it, and refashions it in fresh forms, absolutely indifferent to fact', to what is true or natural or real (20).

Wilde's theory of artistic process can also serve as a starting point in articulating a theory of digital mimesis; of understanding the relationship(s) between the original and its digital manifestation(s), as well as the relationship between and amongst digital surrogates. Moreover, it can be taken as a framework for exploring the complex and shifting relationship between a digitally-presented hyper-reality and material reality.

As has been argued elsewhere, art has never been a mere mirror up to nature, but in no other medium has it been so easy to create a simulacra of reality; a 'heterocosm', simultaneously simulating the familiar while deconstructing it. While the mimetic effect of visualizations, simulations, and virtual reality inherit a set of conventions between an audience and its expectations of a work, these conventions are ultimately unstable, shifting as the technology, and our expectations of it, change.

Digital representations of three dimensional objects, necessarily, lose their corporeality, becoming two-dimensional artifacts[1] engaged with through the mediating presence of an electronic viewing device (a computer monitor, a mobile phone, an e-book). What we engage with, however, are only representations of digital corporeality: what we see are manifestations of the underlying code, much as the prisoners in Plato's allegory of the cave saw only shadows cast on the wall. What we engage with is in fact, not the digital object, but a representation of it.

Johanna Drucker in 'Digital Ontologies: The Ideality of Form in/and code Storage – or – Can Graphesis Challenge Mathesis?' posits that although throughout the Western history, images have been charged with being essentially deceptive or illusionary, the algorithmically-generated code of digital images may, in fact, be a perfect representation of an object; a representation which is not tainted through display or representation. On the other hand, without the representation of the code, the image exists outside our ability to perceive it. In traditional discussions of mimesis the thing being represented typically reflects, however distorted the lens, the represented; the essence of the represented recognizable in the simulacra. With digital media, however, paradoxically, to see beyond the surface of the material world, objects are transmuted into a series of electric currents represented to the computer as binary code. What is being encoded is the object as it never existed, a simulation or hyper-realization.

The intention of a simulation may be to represent an object as it never existed in the material world reflecting our theories and beliefs about it. Digital imagery may be used, for example, to make visible the characters of a manuscript which are no longer perceptible to the human eye. What is represented is not the manuscript as it existed before the damage occurred, nor the manuscript as it exists today: it is not the shadows on the cave wall, nor the reality which casts those shadows, but a hyper-reality which exists between these worlds

As more of our cultural heritage is represented in digital form, the artifacts that people engage with are the simulations without reference to the originals. These disembodied objects exist outside time and space in a way that material objects do not. Digital objects to not decay due to the ravages of time or environment (although digital objects may be rendered useless by our not having the proper hardware and software to read it). Our display paradigms privilege certain readings of these objects; they are surrounded by metadata, typically, if part of a library's holdings, Library of Congress Subject Headings which categorize and group the known world according to a Victorian perception of the universe. Images are not represented to scale, so a map that is 3x2 feet appears the same size as one that is 8x10 inches. Our search engines reduce hundreds, thousands, even millions of objects to a text string displayed ten to a page, or a table populated by 40 2x2 inch thumbnails. This homogenization of results further decontextualize digital simulacra. These deconstructions of the object's material existence reframe the relationship between the perceived and the perceiver, refashioning it, as Wilde writes, 'absolutely indifferent to fact' of what is true or natural or real (20).

This paper will thus explore mimesis from two distinct, but not unrelated aspects of digital technology. The first part will explore the relationship between digital surrogates and their analogue counterparts; how familiar terms like object, imitation,

copy, original function in the digital realm; what is lost and gained in the transfer to the digital when the materiality of a three-dimensional object is transmuted into a two-dimensional plane; the concept of 'trusted digital objects': digital files that will live on when we, and the objects they were created from, no longer exist; the notion that a digital representation may be more appropriately termed a simulacral identity, reflecting, not the object itself, but our beliefs and conventions about it. The second part will explore mimesis from the viewpoint of digital representations as conscious fashionings of hyper-reality or in Wildean terms, employing the unreal and non-existent to recreate the material world in unexpected, fresh, or subversive ways.

Bibliography

Aarseth, Espen J. *Cybertext: Perspectives on Ergodic Literature.* Baltimore, MD: Johns Hopkins University Press, 1997.

Baudrillard, Jean. *Simulations.* New York: Semiotext[e], 1983.

Davis, Michael. *Poetry of Philosophy: On Aristotle's "Poetics"*. South Bend, IN: St Augustine's Press, 1999.

Drucker, Johanna. "Digital Ontologies: The Ideality of Form in/and Code Storage – or –Can Graphesis Challenge Mathesis?" *Leonardo, the International Society for the Arts, Sciences and Technology* 34.2 (2001): 141-145.

Hales, Katherine, N. *Writing Machines.* Cambridge, MA: MIT Press, 2002.

Halliwell, Stephen. *The Aesthetics of Mimesis: Ancient Texts and Modern Problems.* Princeton: Princeton University Press, 2002.

Levy, David M. *Authenticity in a Digital Environment.* CLIR, 2000. <http://www.clir.org/pubs/reports/pub92/levy.html>

Potolsky, Matthew. *Mimesis: The New Critical Idiom.* New York: Routledge, 2006.

San Segundo, Rosa. "A New Conception of Representation of Knowledge." *Knowledge Organization.* 2004. 106-111.

Wilde, Oscar. *The Decay of Lying: An Observation.* Cork: CELT: Corpus of Electronic Texts, 1999. Accessed 2006-09-06. <http://www.ucc.ie/celt/published/E800003-009/>

1. This is true even when the computer emulates three-dimensional space, such as utilizing software to be able to view 380° of a sculpture, or using virtual reality software to emulate perspective.

BFM Old French Text Corpus: Current State and Prospective Developments

Alexei Lavrentiev *(Alexei.Lavrentev@ens-lsh.fr)*

Ecole Normale Supérieure Lettres et Sciences humaines

The BFM (*Base de Français Médiéval*) Old French Corpus was founded in 1989 by Prof. Ch. Marchello-Nizia, and its compilation continues. Céline Guillot has been the project leader since 2006. At present, the main corpus, BFM1, includes 74 complete Old and Middle French texts (approx. 3 000 000 words).

The texts included in the BFM cover a considerable geographic area and an extensive time span, with texts from the 9th century (including the first known French text, the *Serments de Strasbourg*) to the end of the 15th century. Both verse and prose texts are represented, as well as different domains and genres.

The initial aim of the project was to provide scholars with reliable text data on the oldest period of the history of the French language. It was meant to complement the Middle French Text database (14th and 15th centuries) and FRANTEXT database (16th – 21st centuries), both developed in Nancy by the ATILF laboratory. However a number of Middle French texts have been added to the database for different projects.

By now, about 50 theses and a significant number of research books and articles have been prepared with the use of the BFM.

All texts in the main corpus are digitized critical editions. The choice to use editions, and not manuscript transcriptions, was made in order to create an extensive corpus in a relatively short period of time: transcribing a manuscript requires much more time and funds than digitizing a modern edition. However, a subcorpus of manuscript transcriptions is being developed. One of its aims is to provide a precise evaluation of the reliability of critical editions as a source of data for different kinds of linguistic research. In general we argue that using editions or manuscripts are complementary and not mutually exclusive approaches to creating old text corpora (cf. Heiden & Lavrentiev 2004).

All BFM texts are XML encoded with the tags recommended by the TEI. However some of them contain only a very limited markup (headers with some metadata and page breaks or lines

in verse for reference purposes). A richer TEI encoding is applied to a little more than a half of the BFM texts. Its principles are described in (Heiden & Guillot 2002) and a complete description is available at the BFM website.

The BFM texts are not directly accessible to users. They can be searched by means of precise queries (e.g., discrete lexical items, word and phrase concordances, co-occurrences, statistical analyses, etc.) via Weblex search and analysis engine (using CQP query language).

At present, several directions have been chosen for the development of the BFM.

The first direction is concerned with the elaboration of precise text typology, which is necessary to evaluate the representativeness of a text database, a crucial question for all corpus studies. The texts in the BFM are characterized by a number of variables, such as the date, the region, the author, the domain, the genre, etc. The definition of almost each variable is in fact connected with a number of methodological and technical issues. The date can be for instance that of the original text composition or that of the manuscript. Author's age can also be a factor to take into consideration. Only approximate dating is available for many texts and manuscripts.

Domain and genre are the main variables that contribute to characterizing the reprensentativeness of a corpus (cf. Lee 2001). It is however extremely difficult to set a unified genre taxonomy valid over a number of centuries. Many texts belong simultaneously to several domains (e.g. religious and literary, historical and literary).

To deal with this complexity, most of the variables have been encoded by means of multiple fields in the metadata database. Depending on the nature of a query, these can be used in different ways. If the aim of the query is to get a list of texts corresponding to certain criteria, "informal" fields with keyword value type can be used. If the aim is to "cut" the corpus for some kind of contrastive analysis, a unique value is selected for each text on the basis of a formal procedure. If the aim is to place a text in a multidimensional typological environment, it is necessary to model the relations between the different values of a variable and to quantify these values.

The work on text typology in the BFM is in progress, and we will present its current state by the time of the conference.

Another direction in the BFM development is its linguistic annotation. A few texts have been morphologically tagged in a semi-manual mode (with SATO software developed by François Daoust at the University of Québec in Montréal). A complete automatic morphological annotation optimized on the basis of text typology is currently envisaged. Annotation of particular linguistic features (e.g. semantic features of demonstrative adjectives) is conducted in the framework of related linguistic research projects.

The development of the BFM is closely related to that of Weblex. A completely new platform making possible personalized corpus creation and online annotation is under construction.

An important effort is being made to work out consensual text encoding and description procedures with the other projects of diachronic French corpora. The TEI Recommendations are in fact too general to ensure real corpora interoperability. A Consortium for Medieval French Corpora (CCFM) was created in 2004 with the purpose of consolidating efforts of different projects. The BFM team participates in this consortium along with *Laboratoire de français ancien* (University of Ottawa), DMF team (ATILF laboratory, Nancy, France), University of Stuttgart (Germany), University of Zürich (Switzerland), Anglo-Norman Dictionary project (UK) and others.

Web

BFM: <http://bfm.ens-lsh.fr>
Weblex: <http://weblex.ens-lsh.fr/wlx/>

Bibliography

Heiden, S., and C. Guillot. "Capitalisation des savoirs par le web: une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval." *Ancien et moyen français sur le Web, enjeux méthodologiques et analyse du discours.* Ed. Pierre Kunstmann, France Martineau and Danielle Forget. Ottawa: Les éditions David, 2002. 77-92.

Heiden, S., and A. Lavrentiev. "Ressources électroniques pour l'étude des textes médiévaux: approches et outils." *Revue française de linguistique appliquée* IX.1 (2004): 99-118.

Lee, David Y. W. "Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle." *Language Learning & Technology* 5.3 (2001): 37-72.

Exploring New Worlds in Old Texts: Text Encoding Projects for the Undergraduate Study of Spanish American Colonial Literature

Domingo Ledezma
(dledezma@wheatoncollege.edu)
Wheaton College
Phoebe Stinson
(stinson_phoebe@wheatoncollege.edu)
Wheaton College
Scott Hamlin *(hamlin_scotte@wheatonma.edu)*
Wheaton College

Teaching Spanish colonial literature to undergraduate students is a difficult pedagogical task, because students in that level don't have an extensive knowledge of the language of the times or the books' historical context. Indeed, graduate classes and even scholars writing in the field rarely grapple with the early modern editions themselves, because of their accessibility and because they can be so difficult to penetrate. A scholarly collection of essays by Danny Anderson and Jill S. Kuhnheim, that discusses strategies for teaching this material, encourages an approach to these texts based on cultural theory and literary criticism rather than on a close examination of the texts themselves. At Wheaton College (Norton, MA), on the contrary, we have spent the last two years experimenting with a different approach: students' direct involvement with a digital portion of a colonial text (supported by digital cartographic evidence from the period, and mediated by the faculty's suggestions) enhances their comprehension and interest for the actual original book.

In three literature classes undergraduate students have contributed to the creation of a digital critical edition of a 16th century Spanish rare book: *Libro de los Infortunios y Naufragios (Book of the Misfortunes and Shipwrecks)* by Gonzalo Fernández de Oviedo, partially published in Seville in 1535, as part of the *La Historia General de las Indias (General History of the West Indies)*. To date Fernandez de Oviedo's shipwreck narrative has received only scant scholarly attention by literary critics because of the rarity of the book and accessibility to the XVI century printed edition (Kohut).

The class goal is to produce a digital edition of Fernandez de Oviedo's book. The process comprises a variety of steps, aimed to direct students' attention to the texts themselves, their structure, the meaning of words and phrases, and even of individual letters. Students first learn to use accurate transcription techniques and electronic textual editing practices, and then move on to the text encoding based on an XML encoding scheme developed by the Text Encoding Initiative (TEI). We have found that using the TEI encoding is an effective pedagogical tool because, as Allen Renear notes in a article, the standard can "*improve* our ability to describe textual features…. The TEI Guidelines represent an elucidation of current practices, methods, and concepts that open the way to new methods of analysis, new understandings, and new possibilities for representation and communication" (235).

The TEI encoding project functions as a pedagogical tool to study an Early Modern Spanish text; it helps students become more knowledgeable regarding the primary text and the context in which it was written. By using TEI students focus their attention on three main levels: linguistic, geographical and historical. While working directly with a text with the purpose of encoding one of its portions, students have a chance to learn across multiple disciplines, e.g. learning about the ancient Spanish denomination of a place, and creating a link to its location on a map of the times.

The encoding projects are divided into several stages. First, students use text editing software to create a transcription from digital images of original print editions – a process that involves deciphering difficult 16th century typography and finding codes to represent the many characters that are no longer used in contemporary Spanish. After completing the transcription, the students employ a rigorous, descriptive "tagging" process, using the TEI XML encoding scheme. They begin by marking the structural parts of the text – where each chapter begins and ends, each section heading, each paragraph, and so on. They then use TEI encoding to tag historical people names, places, and unfamiliar or archaic vocabulary in the text. And as a final stage of the project, the students perform appropriate research about their texts, and using TEI, define all of the tagged people, places, and vocabulary – essentially providing electronic footnotes to the digital editions of the text. During this stage of the project, students also work with scans of original maps from the period – locating on the maps many places they have tagged, and linking segments of text in their documents to the scans.

This process presents many pedagogical advantages. Students are extremely motivated by projects like this: they work so closely with the text and end up creating their own annotated edition, thus feeling a sense of ownership of the documents.

Many students are excited that this project reaches beyond typical Humanities class work and that they see the results of their hard work very quickly published on the World Wide Web. The students' digital editions of the texts also help preserve and eventually widen the distribution of out of print texts. And finally, this project introduces an academic rigor in studying this literature, which shows in the accuracy taken to encode and validate the text encoding. Within the last few years, Wheaton has been instituting a new curriculum that emphasizes that students should gain a breadth and depth of knowledge through their course work. The process of creating a digital edition of texts like these ensures that students have an in depth experience with a text unlike most other experiences they have had in the humanities. This approach to a text encourages them to apply a systematic, almost scientific approach to humanities scholarship. This work, therefore, is well in line with the goals of many of digital humanities scholarship. A goal, which Susan Hockey expressed in a comprehensive article about the history of the Digital Humanities: “to bring the rigor and systematic unambiguous procedural methodologies characteristic of the sciences to address problems within the humanities that had hitherto been most often treated in a serendipitous fashion” (3).

Our experience teaching with TEI at Wheaton has been very positive: students learn to understand and appreciate the original rare book, only once they have had their experience with creating the digitally enhanced version of it. At that point they’re ready to visit a prestigious library like the John Carter Brown Library at Brown University, and look at the book and at the maps of the period. Undergraduates benefit from the text encoding experience in many ways: better understanding of the topic of study, better understanding of the foreign language, and interest, or at least curiosity, for the rare book itself. Assessment on student learning has been very positive: students retain a lot more than with traditional teaching, and can sometimes also spot irregularities and even fake news in the texts they study. The load on the faculty member and on the library liaison is very high, but it pays off with the students’ satisfaction.

Bibliography

Anderson, Danny J., and Jill S. Kuhnheim. *Cultural Studies in the Curriculum: Teaching Latin America.* New York: Modern Language Association of America, 2003.

Burnard, Lou, and Syd Bauman, eds. *TEI P5:Guidelines for Electronic Text Encoding and Interchange.* TEI Consortium, 2006. <http://www.tei-c.org/release/doc/tei-p5-doc/html/>

Burnard, Lou, Katherine O'Brien O'Keeffe, and John Unsworth, eds. *Electronic Textual Editing.* New York: Modern Language Association of America, 2006.

Cohen, Daniel J., and Roy Rosenzweig. *Digital History : A Guide to Gathering, Preserving, and Presenting the Past on the Web.* Philadlephia, PA: University of Pennsylvania Press, 2006.

Finneran, Richard J., ed. *The Literary Text in the Digital Age.* Ann Arbor, MI: University of Michigan Press, 1996.

Finneran, Richard J., ed. *The Literary Text in the Digital Age.* Ann Arbor, MI: University of Michigan Press, 1996.

Hockey, Susan. "The History of Humanities Computing." *A Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Malden, MA: Blackwell Publishing Ltd, 2004. 3-19.

Kohut, Karl. "Fernández De Oviedo: Historiografía e Ideología." *Boletin de la Real Academia Española* 73.259 (1993): 367-82.

Renear, Allen H. "Text Encoding." *A Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Malden, MA: Blackwell Publishing, 2004. 218-239.

Sperberg-McQueen, C. M., and Lou Burnard, eds. *P4: Guidelines for Electronic Text Encoding and Interchange.* Oxford, UK: TEI Consortium, 2002.

The Electronic Text Center, University of Virginia. *The Etext Center Introduction to TEI and Guide to Document Preparation.* . <http://etext.lib.virginia.edu/standards/tei/uvatei.html>

Women's Writer's Project. *WWP Training Materials.* . <http://www.wwp.brown.edu/encoding/training/index.html>

Human Centered Analysis and Visualisation Tools for the Blogosphere

Xavier Llorà *(xllora@uiuc.edu)*
National Center for Supercomputing Applications
University of Illinois Urbana-Champaign
Noriko Imafuji Yasui *(niyasui@uiuc.edu)*
Industrial and Enterprise System Engineering
University of Illinois Urbana-Champaign
Michael Welge *(mwelge@uiuc.edu)*
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
David E. Goldberg *(deg@uiuc.edu)*
Industrial and Enterprise System Engineering
University of Illinois Urbana-Champaign

1 Motivation

Blogging has become a new and disruptive communication medium. Blogs have changed the way people and organizations express, interact, and—quite unforeseen—exercise influence. David R. Ellis' film *Snakes on a Plane* (2006) starred by Samuel L. Jackson became the first movie to incorporate materials suggested by bloggers long before the movie finished filming. A social mass of blog-based fans influenced the Hollywood creation providing ideas about plots and scenes that finally made it into the released movie. The digital nature of the blog media provides access to an always-expanding corpus of information. It would take more than a lifetime to read all the available blogs necessary to answer questions such as what were the more relevant plots suggested or what key concepts were managed by bloggers in their ideas. However, human-centered analysis and visualization techniques may help users navigate such enormous corpus. This paper presents how human-centered analysis and visualization techniques help identifying relevant post portions and visualizing concept relations in the blogosphere—Google blogs in particular are used for illustrative purposes.

The rest of the paper is structured as follows. Section 2 presents a brief overview of the techniques and visualizations proposed to track the blogsphere. We describe in section 3 how such tools can be applied to track the blogosphere. Finally, we present some conclusions and further research directions in section 4.

2 Snakes, Bloggers, and Human-Innovation

Tracking the blogosphere requires at least (1) gathering blog posts and (2) storing them in a structured metadata store before any analysis and visualization can take place. Blogs rely on syndication feeds, usually incarnating in the form of RSS or Atom feed—both based on XML (Miller, 2001; AtomEnabled, 2006). The first step is to properly process blog feeds by retrieving, annotating, and storing the posts' contents in the feeds for later analysis. Our approach stores the posts in a RDF-based (Shadbolt, 2006) metadata store Mulgara (Gearon, 2006) waiting to be analyzed. Then, we use the extracted text as the input of three different analysis and visualization techniques. It is important to mention here that our approach is based on statistics instead of more traditional approaches based on natural language processing—from which we may benefit in future stages.

2.1 BITS: Getting the relevant terms and excerpts of a post

BITS (blog induced topic selection) is a ranking algorithm for words and sentences in a blog. Higher ranked words may be regarded as main topics used in a blog. Similarly, higher ranked sentences express how key concepts are used in the posts. BITS is inspired by HITS (hypertext induced topic selection) algorithm proposed by Kleinberg (1999). BITS ranking is based on mutually reinforcing relationship between sentences and words: important sentences include many important words and important words are included by many important sentences. The rankings are obtained by an iterative calculation further details can be found elsewhere (Kleinberg, 1999). Each iteration we update the score of each sentence using the sum of scores of all the words of the sentence; we also update the score of words using the sum of scores of all the sentences containing the word.

This mutually recursive calculation provides two important outputs: (1) the ranking of relevant words for a blog, and (2) the ranking of relevant sentences. The ranking of words can be regarded as a summarization of the topics discussed on a given blog. On the other hand, we regard the ranking of sentences as an excerpt extraction technique capable of providing relevant excerpts of a blog and, hence, a summarization tool.

2.3 ISNP: Modelling posts elements

The text contained in a post can be turned into a n-dimensional vector of features using text mining techniques (Weiss, Indurkhya, Zhang, &Damerau, 2006). Each feature is a word in a blog post once stop words are removed. Each vector entry represents a frequency measure for the a given word—TFIDF in our particular case (Weiss, Indurkhya, Zhang, & Damerau, 2006). This simple transformation enables the usage of machine learning techniques as tools for exploring and understanding the processed blog posts. ISNP (Identifying Self/Non-self Post) is an algorithm and visualization technique to create predictive models of posts on a given blog. ISNP uses the postbased feature vectors to learn models that describe and predicts what post belong to a feed. In particular ISNP induce linear models based on support-vector machines (Vapnik, 1999; Cristianini & Shawe-Taylor, 2000; Shawe-Taylor &Cristianini, 2004). Once the models are learned, we can use them to: (1) predict pertinence to a feed given a blog, (2) compare multiple feeds to measure degrees of topic overlapping, and (3) visualize the key elements that identify self in a post.

The proposed visualization based on ISNP results allows the analyst to quickly distinguish the main topics that characterize a feed, and also obtain a measure of the existing overlap between feeds from different posts. The visualization presents a polar arrangement of the terms that distinguish self and non-self blog feeds and the strength of each them—see Figure 1. ISNP also provides another visualization of how topics change as new posts are added to the blogs feed stream by displaying sliding windows of the TFIDF values across the sequences of post of a blog—see Figure 2.

2.3 KeyGraph: Visualizing concept relations

When applied to blogs, KeyGraph (Ohsawa, Benson, & Yachida, 1998) is a chance discovery technique (Ohsawa & McBurney, 2003) which provide a visual map of the contest of the posts of a blog feed. A KeyGraph is a graph where nodes are words on the blog posts and links indicate co-ocurrence of words in sentences. KeyGraph has been widely used as tools to support human innovation and creativity in on-line scenarios (Llor`a, Goldberg, Ohsawa, Matsumura, Washida, Tamura, Masataka, Welge, Auvil, Searsmith, Ohnishi, & Chao, 2006). KeyGraph starts computing high-frequency terms and high-frequency links among them given the sentences of a blog. Then, relevant low frequency terms (key terms) and links (key links) are identified. A key terms and key links bridge high frequency clusters together, flagging interesting transitions between the concepts described by those clusters. Finally, ranking high frequency and key terms based on the connectivity degree allows the KeyGraph to identify keywords.

KeyGraph visualization represents concepts and their relations as visual maps, favoring humanreflection. Moreover, it provides a simple exploratory method to evaluate bridges between concepts, fundamental building blocks of innovation and creativity. KeyGraphs are usually presented nodes and links using three colors: grey to identify high frequency terms and links, red to display key terms and links, and green border nodes to identify keywords—as shown in Figure 3.

3. Tracking the Google Blog

To illustrate the capabilities of BITS, ISNP, and KeyGraphs we tracked the Google Blog (<http: //googleblog.blogspot.com/>) from November 10th to November 14th. A detailed description of the methodology and results is beyond the scope of this paper and can be found elsewhere (Llor`a, Yasui, Welge, &Goldberg, 2007). However, we present some illustrative examples of the blog analysis and visualization techniques proposed in this paper. Unless noted otherwise, the results described below present the analysis of the post entitled *"Old world meets new on Google Earth"*[1]

- *BITS.* The BITS ranking provide the following terms as relevant: *map*, *earth*, *historic*, *explore*, *old*, *world*, *tool*, *and cartography.* The more relevant and descriptive sentence of the post was: *"I was able to explore and fly around the old maps and use the transparency slider to compare the old world and the new; as I did this, I thought to myself that this is the perfect marriage of historic cartographic masterpieces with the innovative contemporary software tools of Google."* After reading the complete post not reproduced due to its length it became clear that the terms provided by BITS were an accurate description of the topics discussed in the post. Moreover, the extracted excerpts by BITS acted as a relevant summary of the posts in the blog.
- *ISNP.* ISNP was used to learn models that uniquely identify posts. The linear model based on support-vector machines was able to accurately distinguish between the ten post of the feed during the period of observation. Moreover, the visualization of such models—see Figure 1— correctly identified topic overlapping with two other posts talking about Google Earth. ISNP also provided a simple visualization of how term relevance changed through time, identifying recurrent topics—Figure 2 displays recurrent topics as wide areas on the accumulated vertical axis graph.
- *KeyGraphs*Finally, the last analysis and visualization tool KeyGraph provided a clear map of the concepts managed in the overlapping posts *"Old world meets new on Google Earth"* and *"Know where you are"*[2], as well as their bridging relations see Figure 3. KeyGraph clearly visualize the two main discourse clusters provided by the

overlapping posts, and also made explicit the connection between them.

4. Conclusions and Further Work

This paper has presented how human-centered analysis and visualization techniques used to support innovation and creativity can also help to identify relevant post portions and to visualize concept relations in the blogosphere. BITS, ISNP, and KeyGraphs were introduced and used to analyze the posts on the Google Blog for illustrative purposes. The proposed techniques showed how humancentered techniques can easily assist tracking the blogosphere for relevant information, concepts, and relations, filtering the amount of information that the analyst need to review by providing meaningful summaries and visualizations.

Acknowledgments

We would like to thank the Automated Learning Group at the National Center for Supercomputing Applications for their friendship and support while hosting this joint collaboration. This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant F49620-03-1-0129, and the National Science Foundation under grant IIS-02-09199. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

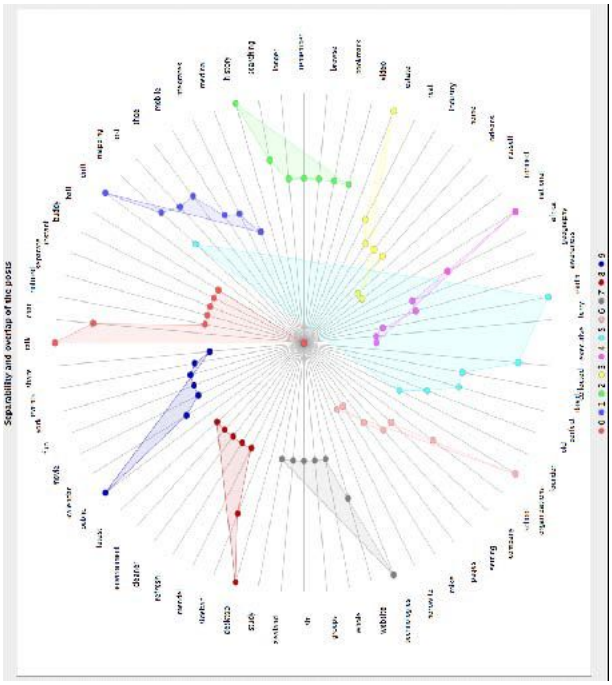


Figure 1: Radial map of the key terms involved in the ISNP models for each of the posts. Different posts are displayed in different colors. The area provided by the measure of relevance of the terms provide a qualitative measure of model overlapping for the different post. Post number 5 corresponds to the analyzed "Old world meets new on Google Earth".

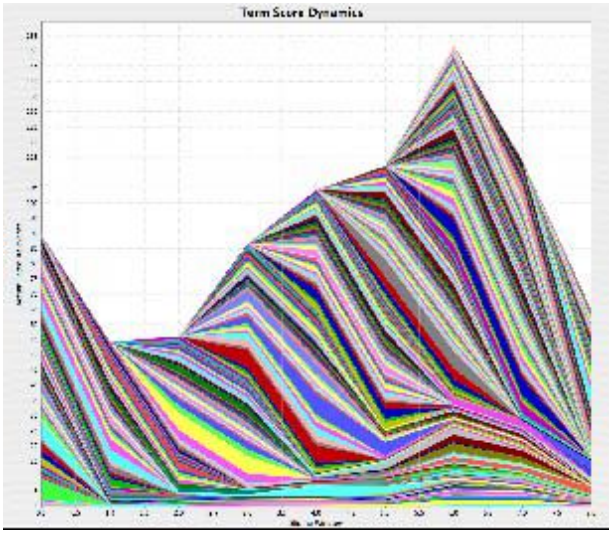


Figure 2: ISNP visualization of term dynamics across the different post.

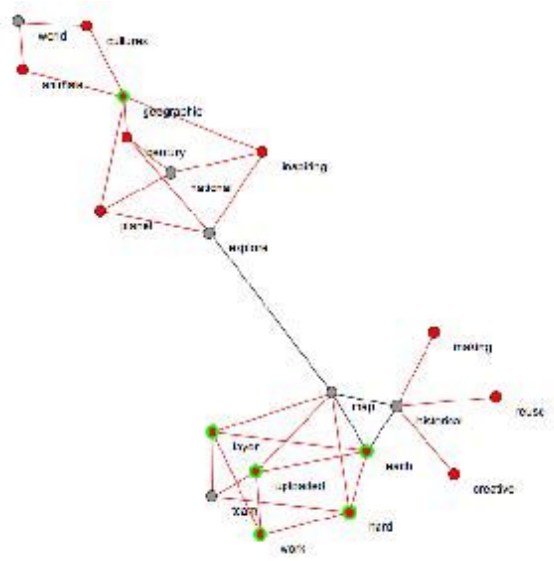


Figure 3: Visual map of the concepts involved in the two overlapping posts "Old world meets new on Google Earth" and "Know where you are". KeyGraph clearly visualize the two main discourse clusters provided by the overlapping post, and also makes explicit the connection between them.

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 the Air Force Office of Scientific Research, the Technology Research, Education, and Commercialization Center, the Office of Naval Research, the National Science Foundation, or the U.S. Government.

1. <http://googleblog.blogspot.com/2006/11/old-world-meets-new-on-google-earth.html>
2. <http://googleblog.blogspot.com/2006/11/know-where-you-are.html>.

Bibliography

AtomEnabled, A. *What is atom?* . <http://www.atomenabled.org/>

Cristianini, N., and J. Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge Press, 1997.

Gearon, P. *Mulgara Metadata Store* . <http://www.mulgara.org/>

Kleinberg, J. "Authoritative Sources in a Hyperlinked Environment." *Journal of the ACM* 46.1 (1993): 604-632.

Llorà, Xavier, M. Welge, N. I. Yasui, and D. E. Goldberg. "Analyzing Trends in the Blogosphere Using Human-Centered Analysis and Visualization Tools." *Proceedings of the International Conference of Weblogs and Social Media.* in press.

Llorà, Xavier, et al. "Innovation and Creativity Support Via Chance Discovery, Genetic Algorithms, and Data Mining." *New Mathematics and Natural Computation* 2.1 (2006): 85-100.

Miller, E. *W3C RSS 1.0 news feed creation how-to* . 2001. <www.w3.org/2001/10/glance/doc/howto>

Ohsawa, Y., B.E. Benson, and M. Yachida. "KeyGraph: Automatic Indexing by Cooccurence Graph Based on Building Construction Metaphor." *Proceedings of Advances in Digital Libraries.* 1998. 12-18.

Ohsawa, Y., and P. McBurney. *Chance Discovery.* Springer, 2003.

Shadbolt, N. "The Semantic Web Revistited." *IEEE Inteligent Systems* 21.3 (2006): 96-101.

Shawe-Taylor, J., and N. Cristianini. *Kernel Methods for Patter Analysis.* Cambridge Press, 2004.

Vapnik, V. *The Nature of Statistical Learning Theory.* Springer, 1999.

Vapnik, V. *Kernel methods for patter analysis.* Springer, 1999.

Weiss, S., N. Indurkhya, T. Zhang, and F> Damerau. *Text Mining: Predictive Methods for Analyzing Unstructured Information.* Springer, 2006.

The Digital Museum in the Life of the User

Paul F. Marty *(marty@ci.fsu.edu)*
College of Information
Florida State University

Libraries, archives, museums, and other cultural heritage organizations are creating digital representations of their artifacts and collections at an astounding rate (Institute of Museum and Library Services, 2005). The potential of increased access to digital collections has tremendous implications for digital humanities researchers, and many recent projects have focused on the technical mechanics of digitizing cultural heritage resources. While these considerations are extremely worthwhile, there is an equally important need to explore how people use these digital resources once they become available (Cameron, 2003; Knell, 2003).

This paper explores the role digital museum resources play in the lives of the users of those resources. It presents results from an international survey (administered to nearly 1500 visitors at nine different online museums) that addressed the relationship between digital and physical museum resources in the lives of museum visitors. The survey questions focused on how museum professionals can encourage their visitors to form lifelong relationships with museums: visiting in person when they can, and visiting online when they cannot. The results of the survey help museum researchers and professionals understand the complementary, cyclical relationship that exists between digital and physical museums from a user-centered perspective.

A growing number of researchers are interested in studying how the many different users of museum resources (museum visitors and professionals, information providers and consumers) employ digital museum resources in their everyday lives (White, 2004). Museum professionals today work in a changing environment where information resources are becoming more technically-complex, and where the users of those resources are becoming more information-savvy. Over the past decade, the needs and expectations of museum visitors have become increasingly sophisticated, and museum professionals are increasingly concerned with ensuring that the right information resources are available to all users, inside and outside the museum (Marty, Rayward &Twidale, 2003).

To meet the changing needs of their visitors, museum professionals have dramatically changed the way museum visitors interact with museum resources. Increased access to digital collections has removed many traditional barriers between museums and their visitors, offering new opportunities for interacting with collections and information resources. Douma and Henchman (2000), for example, discuss an online exhibit that allows visitors to digitally "strip away" layers of a painting (Bellini's Feast of the Gods), examining earlier versions using simulated infrared or x-ray lenses. Gillard (2002) explores how the National Museum of American History's HistoryWired project encourages visitors to manipulate a collection of artifacts, uncovering connections between objects along temporal, cultural, and thematic lines.

These changes have led museum professionals to express concerns about the impact new information technologies have on the relationship between museums and their visitors. Some worry whether providing online access to digital museum resources will result in a decrease in visits to physical museums and a corresponding loss of financial revenue (Haley-Goldman & Wadman, 2002). In the process of providing access to digital museum resources, someone usually asks, "if visitors can access these resources over the Internet, will they still come to the museum?" The simple answer to this question is "yes, they will," and a number of recent surveys have shown that online visitors are also physical visitors (Kravchyna & Hastings, 2002). This makes sense; nobody asks, "if people can look at pictures of beaches online, will they still vacation in Florida?" Similarly, the ability to access digital museum resources online should in theory serve as a lure, encouraging potential visitors to come to the physical museum.

Despite this commonsense approach, worries about the relationship between physical and digital museum resources persist. Why? The truth is that these worries have far less to do with financial remuneration than with understanding how the users of digital museum resources perceive the integration of those resources into the sociocultural fabric of their everyday lives. To understand the relationship between digital and physical museums, one must ask, "what role do digital museum resources play in the life of the user of digital museums?" Framing the question in this manner reflects a shift in perspective away from the "user in the life of the museum" to the "museum in the life of the user"—a shift that parallels one that has taken place in the library and information science community over the past few decades.

Very little is known about why users seek digital museum resources or how users integrate digital museum resources into their everyday lives, despite valuable studies that have explored what visitors do at museum websites (Thomas & Carey, 2005). Understanding the relationship between digital museums and their users becomes critically important as more museums offer digital resources online, and the number of visitors to online museums increases to be five to ten times the number of visito.rs to physical museums. If one wants to encourage a situation

where visitors make museums part of their everyday lives and feel connected to museums whether they are physically there or not, there is a need to explore the digital museum in the life of the user.

To meet this need, this study explored the following research questions:

- What is the role of the digital museum in the life of the museum visitor?
- How can museums use museum websites to build stronger relationships with their visitors, before and after a museum visit?
- What needs do digital museum resources meet for museum users outside the museum?
- When and how do people use museum websites before and after visiting museums?
- What do visitors prefer to do on museum websites vs. in the museum, and vice versa?
- How do museum websites influence the visitors' desire to visit the museum?

To answer these questions, the researcher developed an online survey that asked online museum visitors about their use of digital museum resources before and after museum visits, and how they integrate digital and physical museum resources in their everyday lives. The survey was advertised on the websites of nine different museums, including the Fine Arts Museums of San Francisco, the Science Museum of Minnesota, the Australia War Memorial Museum, the Victoria and Albert Museum, and the National Museum of Wildlife Art. From October 2005 to October 2006, 1464 online visitors responded to the survey.

The survey results provide valuable insights into the behavior of visitors to online museums around the world. Certain key findings from the study are summarized in the following list:

- Online museum visitors consider it very important for museums to have a website.
- They are very likely to visit a museum's website before visiting the museum.
- They are likely to use a museum's website to determine whether they want to visit the museum.
- They have strong preferences for what they want to do in the museum vs. what they want to do using the museum's website.
- They are likely to visit the museum's website after visiting a museum.
- After leaving a museum, they expect to be able to find the museum's website easily, and they rely on the museum's website to answer questions about the museum.
- They are likely to establish a relationship where they visit a museum and its website repeatedly, visiting the museum when they can, and its website when they cannot.
- They are likely to visit museum websites in their daily life, independent of planning or returning from a museum visit.

These results indicate that online museum visitors view museums and museum websites as complementary, and that digital museum resources are not likely to replace physical museum resources in the lives of museum visitors. The users of digital museums are constructing, mostly on their own initiative, a complicated relationship between digital and physical museum resources in their own lives. By focusing solely on the user in the life of the museum, one sees only how visitors use resources there, not how they make use of all museum resources, digital and physical, in the museum and online, in their own lives. Studies of the digital museum in the life of the user, therefore, are more likely to paint an actual picture of the complex interactions the users of museum resources experience at the boundary of physical and digital museums.

This study explores the broader implications of access to digital culture by addressing the relationship between physical museum objects and digital information resources. Much of the world's cultural resources are located in small museums, historical societies, and community heritage associations—organizations that are just now digitizing their collections. Without a solid understanding of how people use digital museum resources in their everyday lives, what good does it do those institutions to create those same resources? If one is to create a sustainable future for the digital humanities, one must improve the overall understanding of how individuals use digital information resources in ways that augment their appreciation and understanding of physical artifacts and cultural heritage worldwide.

Bibliography

Cameron, F. "Digital Futures I: Museum Collections, Digital Technologies, and the Cultural Construction of Knowledge." *Curator* 46 (2003): 325-340.

Douma, M., and M. Henchman. "Bringing the Object to the Viewer: Multimedia Techniques for the Scientific Study of Art." *Museums and the Web 2000: Selected Papers from an International Conference.* Ed. D. Bearman and J. Trant. Pittsburgh, PA: Archives &Museum Informatics, 2000. 59-64.

Gillard, P. "Cruising Through History Wired." *Museums and the Web 2000: Selected Papers from an International*

Conference. Ed. D. Bearman and J. Trant. Pittsburgh, PA: Archives &Museum Informatics, 2002.

Haley Goldman, K., and M. Wadman. "There's Something Happening Here, What it is ain't Exactly Clear." *Museums and the Web 2000: Selected Papers from an International Conference*. Ed. D. Bearman and J. Trant. Pittsburgh, PA: Archives &Museum Informatics, 2002.

Institute of Museum and Library Services. *Status of Technology and Digitization in the Nation's Museums and Libraries*. 2005. Accessed 2006-10-31. <http://www.imls.gov/resources/TechDig05/Technology+Digitization.pdf>

Knell, S. "The Shape of Things to Come: Museums in the Technological Landscape." *Museum and Society* 1.3 (2003): 132-146.

Kravchyna, V., and S. Hastings. "Informational Value of Museum Web Sites." *First Monday* 7.2 (2002). Accessed 2006-10-31. <http://firstmonday.org/issues/issue7_2/kravchyna>

Marty, P.F., W.B. Rayward, and M. Twidale. "Museum Informatics." *Annual Review of Information Science and Technology* 37 (2003): 259-294.

Thomas, W.A., and S. Carey. "Actual/Virtual Visits: What are the Links? ." *Museums and the Web 2005*. Ed. D. Bearman and J. Trant. Toronto, CA: Archives &Museum Informatics, 2005.

White, L. "Museum Informatics: Collections, People, Access, and Use." *Bulletin of the American Society for Information Science and Technology* 30.5 (2004): 9-10.

Digital Editing, Infrastructure Obstacles, and the World of Virtual Appliances

Jarom Lyle McDonald
(jarom_mcdonald@byu.edu)
Brigham Young University

A challenge that is often too daunting for new collaborative digital editing projects is that of production and publication infrastructure. The TEI Consortium has done a wonderful job of providing a wealth of resources to speak to this need for infrastructure support; with the online guidelines, samples, and stylesheets, to the Roma customization generator, to the (underutilized) TEI wiki, a project in the planning stages has the ability to gather more than enough information and advice for developing a solid infrastructure for software encoding, database storage, web delivery.

Unfortunately, information and advice often isn't enough, and projects which don't have access to sufficient technical skills sometimes cannot even get their database installed, their web server up and running, or their transformations working properly. This lack of infrastructure support often leaves promising projects floundering and potential digital humanities scholars soured on the ability of technology to deliver on its revolutionary promises.

There are some viable products that have become available in the past few years specifically designed to alleviate this problem of a general lack of infrastructure support. For example, <teiPublisher> a Publishing system built on the eXist database and the Lucene search engine, describes itself as "an extensible, modular and configurable xml-based repository . . . designed to provide administrative tools to help repository managers with limited technical knowledge manage their repositories" (http://teipublisher.sourceforge.net). More ambitiously, Sebastian Rahtz has created a collection of Linux packages at http://tei.oucs.ox.ac.uk/teideb/ which provide instant TEI-informed installation of such products as eXist, the Cocoon publishing system, the Saxon XSL transformation engine, and so forth. However such products are not in widespread use, perhaps because those who need them the most are even struggling with more basic infrastructure problems such as getting a web server installed.

Enter the appearance of virtual appliances. The world of hardware virtualization has exploded over the past few years;

IT professionals from a wide variety of fields have seen the benefits of having a single piece of hardware run multiple "virtual machines," each appearing to be a real machine to the outside networked world. Virtual appliances are small imprint virtual machines, each designed for a different job, that can reside on a physical machine and perform some sort of task in relation to other virtual appliances (or physical machines) that it is networked to. As appliances rather than software packages, they are designed to be nearly functional out of the box, with some basic configuration usually performed through a web interface but with most of the configuration pre-installed, pre-set, and pre-tested. Virtual machines and virtual appliances have been successfully used in commercial and educational settings to provide instant infrastructure support, allowing users to concentrate on what their tasks are to be rather than how things work underneath – in other words, well-designed virtual appliances should function as turn-key devices, alleviating the need for a given project to have a dedicated systems administrator.

The purpose of this poster session, then, is to put forth the concept of virtual appliances as an answer for the obstacle that many digital editing projects face. I will create a set of virtual appliances that can be deployed as a group—one will be a web server, one a relational database server, one an XML database server, one an XML transformer, and so forth—that forms the core foundation of an online digital project. Each of the virtual appliances is a configured "machine," with only basic information needing to be set through a web interface to each of the pieces. A project hoping to utilize this group of virtual appliances would perform the following steps:

1. A project will procure a physical machine.
2. The project will install the software to run the virtual machines. In the case of this poster session, this will be VMWare Server (free of charge), as VMWare is one of the leaders in the development of virtual machine technology (see <http://www.vmware.com/vmtn/appliances/faq.html>); however, conceptually, the same type of virtual appliance network could be developed for the Xen virtualization software.
3. The project will download the virtual appliances it chooses onto the physical machine, one of which will be a central control machine.
4. The project will use a web interface to the central control machine to supply the relevant details—IP address, project name, etc.
5. The project will then have an instant network of "servers" that can act as a home for collaborative encoding storage, database management, and web delivery of the materials to be published.

Obviously, there are many issues that my proposal does not seek to solve—projects will still have to worry about creating materials, for example, or setting up a higher-level front end for accessing the materials. But even these problems might, down the road, find a solution in virtual appliance technology; imagine, perhaps, that one of the virtual appliances in this network runs a pre-configured installation of teiPublisher, thus providing a repository management system for project collaborators. Another virtual machine might have nothing on it but Roma and some validators, and act as a localized (yet still networked) project schema management system. A third could have a turn-key wiki for collaborative documentation writing.

By seeking to use virtualization technology and the concept of ready-made, plug-and-play virtual machines, I hope to make the world of digital editing and publishing available to a wider range of scholars and users. This new model of "software" distribution—treating infrastructure as appliances—will eliminate many of the lower-level technical obstacles that prevent too many good ideas from finding a home in the digital world.

A Network Structure of the Synoptic Gospels Employing Clustering Coefficients

Maki Miyake *(mmiyake@lang.osaka-u.ac.jp)*
Osaka University

In this presentation, we propose a graph-theory approach to the systematization of the texts of the Synoptic Gospels. Specifically, we apply a new graph-clustering technique for data processing that utilizes a clustering-coefficient threshold in creating semantic networks for the Gospels. Taking Greek texts (Nestle-Aland, 1979) as the textual source, an adjacency matrix for graphs is created from word co-occurrence data for a given range of the texts. We also develop a web-based application for the dynamic representation of the network structures based on the relationships between words and concepts.

Network representations are useful in modeling natural language semantics (Stevers, & Tenenbaum, 2005). Recently, Dorow (2006) has proposed the notion of curvature (clustering coefficient) as a clustering tool to detect ambiguity and to acquire semantic classes rather than using word frequencies. The clustering coefficient represents the interconnective strength between neighboring nodes in a graph. Following Watts and Strogatz (1998), the clustering coefficient of the node (n) can be defined as: (number of links among n's neighbors)/(N(n)*(N)-1)/2), where N(n) denotes the set of n's neighbors. The coefficient assumes values between 0 and 1.

In processing a text, word pairs are computed by a windowing method (Takayama, Flournoy, Kaufmann, &. Peters). The windowing method provides a relatively simple representation of similarity levels that is suitable for clustering. The technique involves moving a certain sized window over a text to extract all fixed-sized word grams (Vechthomova, Roberston, & Jones, 2003). Word pairings are subsequently made by combining all extracted words. Co-occurrence data is computed with two window sizes that reflect syntactic and semantic considerations. The first size is set at 2 for the nearest co-occurrence, while the second size is set to 20 to collect words both before and after a sentence. Morphological data for the BibleWorks Greek New Testament Morphology (BNM) was used in a stemming process to obtain base word forms from morphological variants. Consequently, 2,688 word occurrences were identified.

In Figure 1, the degree distribution for the word occurrences nodes are plotted in log-log coordinates with showing the best fitting power law distribution (r=1.4). Following Barabasi and Albert (1999), the feature of power-law degree distribution indicates the small-world structure of the semantic network. Figure 2 is a plot for the clustering coefficient as a function of degree. There are 16 words that have a clustering coefficient value of less than 0.1. These are articles, prepositions, pronouns, and conjunctions, which are usually regarded as "noise" words. We use the clustering coefficient value as a threshold in order to eliminate such noise words and to control the datasets for graph clustering. In applying the clustering technique, nine datasets were created with 0.1 increments in the value of the clustering coefficient (from 0.1 to 0.9). The datasets for each window size were converted into adjacency matrices for Recurrent Markov Clustering (RMCL) (Jung, Miyake, & Akama, 2006).

RMCL is an improved form of Markov Clustering (MCL) (van Dongen, 2000). The MCL algorithm is based on random walks for a graph, and its model simulates flow by using the two simple algebraic operations of expansion and inflation on the stochastic transition matrix. The method has been applied to a number of corpora, such as a French synonym dictionary (Gfeller, 2005) and the British National Corpus (Dorow, Widdows, Ling, Eckmann, Danilo, and Moses, 2005). In contrast, RMCL allows for greater control over the sizes of concept domains by modifying graph granularity and the generality of concepts. The recurrent process gets feedback from the state of overlapping clustering before the output of MCL. This reversal procedure is a key feature of RMCL in generating a virtual adjacency matrix for the non overlapping clusters as a resultant state of convergence actually yielded by the MCL process. The resultant downsized matrix provides a simpler graph of the conceptual structures underlying similar words.

Figure 3 shows the transitions in cluster sizes for the MCL and RMCL processes as a function of the clustering coefficient with a window size of 2, while Figure 4 shows similar data with a window size of 20. The term input data refers to the initial adjacency matrices. Figure 4 indicates that while cluster sizes are almost identical for RMCL and MCL at clustering coefficient values of 0.6 or more, RMCL clearly yields smaller clusters at values of 0.5 or less. The MCL clustering results in Figure 4 for the window size of 20 indicate the strong connectedness between nodes, as all nodes are grouped into a single cluster for coefficient values of 0.2 or less.

The results for both the MCL and RMCL processes are implemented as graph network structures in a web application, which has been developed to dynamically represent the relationships between words and how MCL components might correspond to concepts. Figure 5 is a screen shot of the web

application, which uses Apache2 as an http server and Tomcat5 as a JSP and servlet engine. WebMathematica is used to call a Mathematica kernal to calculate the graphs.

In conclusion, the clustering coefficient represents a useful tool for manipulating datasets to eliminate noise words. The RMCL methods allows for the creation of compact semantic networks, which clearly present the relationships between words.

In future works, we will apply different clustering-coefficient thresholds to detect ambiguous words, particularly words with low coefficient-low degree values. We will also evaluate systematizations of the Synoptic Gospels by biblical scholars and explore the effectiveness of our approach in furthering biblical studies.

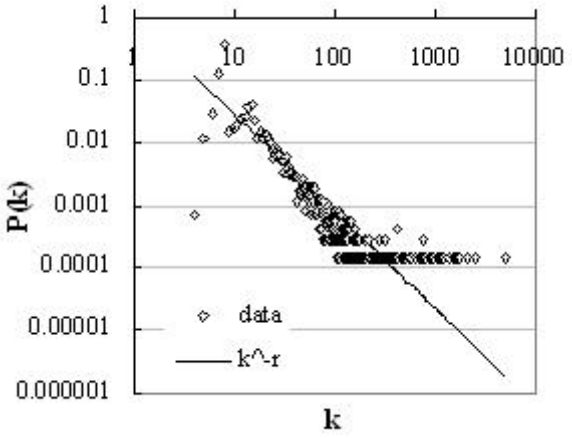


Figure 1: The Degree Distributions

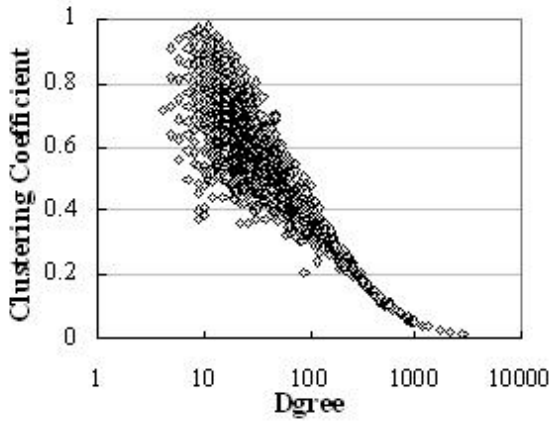


Figure 2: Clustering coefficient plotted as a function of degree

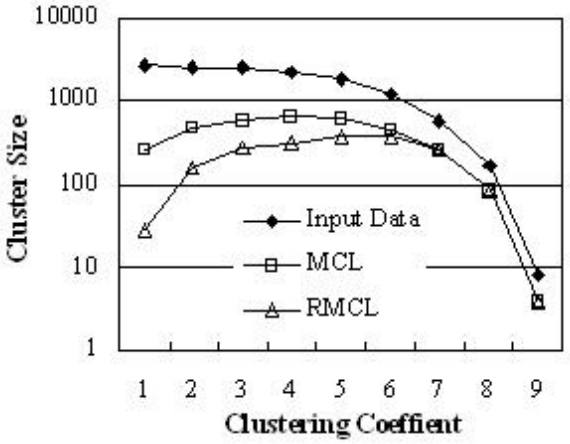


Figure 3: Clustering Coefficient (window size =2)

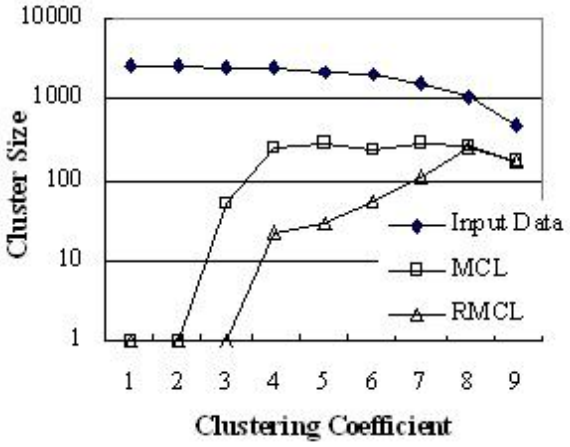


Figure 4: Clustering Coefficient (window size =20)

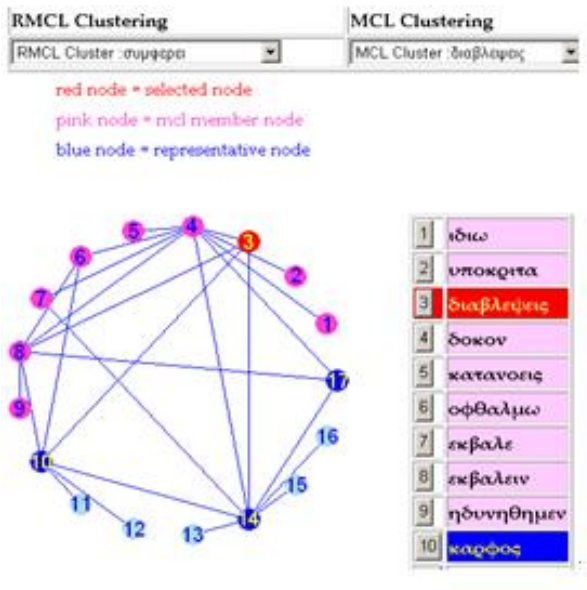


Figure 5: Screen shot of a web application

Bibliography

Baranbasi, A.L., and R. Albert. "Emergence of Scaling in Random Networks." *Science* 286 (1999): 509-512.

Dorow, B., D. Widdows, K. Ling, J. Eckmann, D. Sergi, and E. Moses. "Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination." *MEANING-2005, 2nd Workshop organized by the MEANING Project, February 3rd-4th 2005, Trento, Italy*. Trento,Italy, 2005.

Dorow, K. "A Graph Model for Words and Their Meaning". Doctoral Thesis. University of Stuttgart, 2006. <http://elib.uni-stuttgart.de/opus/volltexte/2007/2985/pdf/diss_27022007.pdf>

Gfeller, D., J. C. Chappelier, and P. De Los Rios. "Synonym Dictionary Improvement through Markov Clustering and Clustering Stability." *International Symposium on Applied Stochastic Models and Data Analysis*. . 106-113.

Jung, J., Maki Miyake, and H. Akama. "Recurrent Markov Cluster (RMCL) & #12288; Algorithm for the Refinement of the Semantic Network, LREC2006lity." *International Conference on Language Resources and Evaluation*. . 1428-1432.

Nestle-Aland. *Novum Testmentum Graece* . 26th edition. Stuttgart: German Bible Society, 1987.

Steyvers, M., and J. B. Tenenbaum. "The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth." *Cognitive Science* 29.1 (2005): 41-78.

Takayama, Y., R. Flournoy, S. Kaufmann, and S. Peters. *Information Mapping: Concept-based Information Retrieval Based on Word Associations*. . <http://www-csli.stanford.edu/semlab-hold/infomap/theory.html>

van Dongen, Stijn Marinus. "Graph Clustering by Flow Simulation". PhD Thesis. University of Utrecht, 2000. <http://igitur-archive.library.uu.nl/dissertations/1895620/inhoud.htm>

Vechthomova, O., S. Robertson, and S. Jones. "Query Expansion With Long-Span Collocate." *Information Retrieval* 6 (2003): 251-273.

Watts, D., and S. Strogatz. "Collective Dynamics of 'Small-World' Networks." *Nature* 393 (1998): 440-442.

Quantitative Data, Formal Analysis. Reflections on 7,000 Titles [British Novels, 1740-1850]

Franco Moretti *(moretti@stanford.edu)*
Center for the Study of the Novel
Stanford University

This opening plenary presentation provided an overview of a century of British literary history as reflected in novelistic titles. Three main issues were investigated: the dramatic shortening of titles, and its possible causes; the structure and significance of extremely short titles; and the ways in which titles allude to specific literary genres.

Roundtable Panel: Modeling and Visualizing Historical Narrative

Ruth Mostern *(rmostern@ucmerced.edu)*
Social Sciences, Humanities and Arts
University of California, Merced
Johanna Drucker *(jrd8e@virginia.edu)*
Media Studies
University of Virginia
Ian Johnson *(johnson@acl.arts.usyd.edu.au)*
Archaeological Computing Laboratory
University of Sydney
Lewis Lancaster *(buddhst@berkeley.edu)*
Electronic Cultural Atlas Initiative
University of California, Berkeley
Bruce Robertson *(broberts@mta.ca)*
Classics
Mount Alison College

The modeling and visualization of temporal phenomena—events and narratives—is an important area of research in the digital humanities that has received relatively little attention to date. There are some exceptions. Matt Jensen's SemTime, Johanna Drucker and Bethany Nowviskie's Temporal Modeling Project, the Center for History and New Media's Timeline Builder, and Bruce Robertson's Historical Event Markup Language are important projects that have been presented to or developed within this community. However, in spite of these exceptions, the limits of work in this area are clear when compared, for instance, with the closely allied area of spatial modeling and historical GIS. In recent years, interactive digital maps have become increasingly prevalent. In addition to the ubiquitous Google Earth, other map resources produced by universities and textbook publishers allow users to pan and zoom, hyperlink to related content, control layers of spatial information, and select temporal ranges. However, history textbooks are filled with timelines as well as maps, and temporal thinking is critical not only to history, but to many other disciplines besides. Nevertheless, in contrast to mapping, the problem of modeling and visualizing time in an interactive and digital environment has received almost no attention.

With the hope of raising the visibility of current efforts and inspiring new research, this panel brings together six individuals working in the area of temporal modeling and interactive timeline development. Our goal is to introduce current developments; to identify interesting and significant areas of development that will make temporal modeling and visualization into a robust field of research, and to discuss models for an interactive database and historical event visualization system.

Some components of a historical event model are well developed. Temporal ontologies such as DAML-Time and ISO 19108 offer essential guidance for formalizing temporal objects. The Historical Event Markup Language and, more recently, the Named Time Period Directory Standard have developed XML models for describing historical events. Creative efforts to improve timeline visualization are surprisingly limited, but there are several noteworthy exemplars. The MIT Media Lab's SIMILE project has an intuitive and flexible interface and allows several timelines at different scales to be manipulated together. However, it implicitly embodies a historical model of individual unambiguous and unrelated events rather than complex narratives, replicating the "tickmarks on a line" model familiar from static textbook and wall chart timelines. By contrast, the Temporal Modeling Project and SemTime are visualization experiments that allow developers to incorporate agency, causality and relationships.

Better handling of relationships is among most important issues for future development of temporal modeling in the digital humanities. Historical events have complex, multiple and perspectivally unique relationships to one another. Historical events and narratives also all have a relationship to spatial information. Events occur somewhere, even when they are global, and the same event (the Neolithic, modernity) may occur at quite different times in different places. Finally, timelines and event databases can be considered as components of holistic information systems. All of these issues will be explored in this roundtable session. Other topics that have not been addressed by previous researchers in temporal modeling include the representation of temporal and temporo-spatial ambiguity and uncertainty, and the specifications for historical event systems and services with multiple users. These topics, too, will feature in our planned roundtable. All of these areas have implications for modeling of both data and systems, and also for very new kinds of visual representation.

The proposed Modeling and Visualizing Historical Narrative Roundtable brings together a number of individuals who are working actively in this area. This panel includes some individuals (Drucker and Robertson) who have been researching temporal modeling and visualization for some time, and others (Johnson, Lancaster, Mostern, all affiliated with the Electronic Cultural Atlas Initiative) who have worked primarily on spatial visualization for history and the humanities and have begun to engage in research on temporality as an extension of that interest.

Ruth Mostern, as panel organizer, will introduce the panel with a discussion of work-to-date on temporal modeling and possible directions for future research. The rest of the panelists will discuss their own areas of development, as follows:

- *Ian Johnson* will discuss developments in the Heurist generic collaborative content creation system and the TimeMap spatial browser to create a model and visualizer for historical events, illustrated through the example of a Silk Road timeline project. He will discuss the underlying data structures, planned developments in the data model used to describe historical events, and research on improved visual methods of entering and viewing the web of relationships between historical events, including timeline visualizations.
- *Lewis Lancaster* will demonstrate the What, Where, When and Who prototype for library catalogue searching and web browsing. In this system, a directory of named events extracted from Library of Congress subject headings serves as the basis for structured searching. Searchers can browse a library catalogue by using a timeline and a map linked to the time period directory and a place-name gazetteer rather than conducting a traditional text-based search. The user may expand the search from a library catalogue to include Wikipedia and other on-line systems. Named events, place names, and personal names are interrelated to allow complex search parameters.
- *Bruce Robertson* will talk about the latest developments for the Historical Event Markup Language. The first of these is an RDF syntax for historical events that makes historical markup of timed, labeled events associated with persons, places and keywords simple for students and others. The other development is the use of OWL web ontology reasoning to express chronological reasoning (i.e. how we know that the battle of Marathon was in 490 BC). This will be associated with RDF reification
- *Johanna Drucker*, an important researcher in this area, regrets that she will not able to attend the conference due to a prior commitment. However, she has offered to produce visualizations based on her recent work to be presented by the conference organizer.

Because of the large number of presenters and the desirability of open dialogue, we are proposing a roundtable format. The panel organizer and each presenter (including Drucker *in absentia*) will make a presentation of 10 minutes. Each presenter will discuss his or her current work and future plans in the area of temporal modeling and visualization. This will allow time for a moderated discussion among the presenters and between the presenters and the audience.

Collex: Facets, Folksonomy, and Fashioning the Remixable web

Bethany Nowviskie *(bethany@virginia.edu)*
Applied Research in Patacriticism
University of Virginia

Collex is an online toolset designed to aid students and scholars working in networked archives and federated repositories of humanities materials: a sophisticated *COLL*ections and *EX*hibits mechanism for the semantic web. It allows users to search, browse, annotate, and tag electronic objects and to repurpose them in illustrated, interlinked essays or exhibits. By saving information about user activity (including the construction of annotated collections and exhibits) as "remixable" metadata, the *Collex* system writes current practice into the scholarly record and permits knowledge discovery based not only on predefined characteristics or "facets" of digital objects, but also on the contexts in which they are placed by a community of scholars. Collex builds on the same semantic web technologies that drive MIT's SIMILE project and social bookmarking systems like *Connotea* and *Zotero*, but it also brings folksonomy tagging to trusted, peer-reviewed scholarly archives and features an integrated publication system. This exhibits-builder is analogous to high-end digital "curation" tools currently affordable only to large institutions like the Smithsonian. *Collex* is free, generalizable, and open source and is presently being implemented in a large-scale pilot project under the auspices of NINES.

Collex is constructed with pragmatic scholarly needs in mind, and under the assumption that "the general field of humanities education and scholarship will not take the use of digital technology seriously until one demonstrates how its tools improve the ways we explore and explain our cultural inheritance – until, that is, they expand our interpretational procedures" (McGann, Radiant *Textuality* xii; my emphasis). *Collex* facilitates primary interpretive gestures of exploration and explanation in a broad and socially-networked manner, and aims to form a locus for further expansion of interpretive methods in digital humanities.

The first formal iteration of *Collex* (released in February 2007) federates more than 60,000 digital objects of 19th-century literature, art, culture and criticism from the most prominent and acclaimed online journals, archives, and repositories in the field. This pilot project forms the core of NINES, the *Networked Infrastructure for Nineteenth-century Electronic Scholarship*, a trans-Atlantic federation of scholars and of peer-reviewed primary and secondary materials constituting a federated collective. Endorsed by the NINES steering committee and under development for the past year at ARP, the University of Virginia's *Applied Research in Patacriticism* lab, *Collex* is both the central clearing-house for NINES and the interpretive hub around which we hope a vital community of scholars and students will coalesce.

Humanists eager to develop new ways to integrate and explore digital works currently lack crucial institutional and technical resources. Even the best models that scholars of (for instance) nineteenth century literature and culture now follow and imitate — the *Whitman Archive*, *Romantic Circles*, the *Rossetti Archive*, the *William Blake Archive* — are stand-alone projects that can only be loosely integrated through web browsers, even when shared through OAI protocols. As a consequence, what you see now on the web is what you get: an agglomeration of sites and projects whose content is atomized and whose scholarly and educational value is indeterminate. While it is possible for tech-savvy scholars, using ad-hoc tools and methods, to produce and distribute annotated, re-organized, or selected versions of existing online resources, they presently lack coordination within a peer-reviewed digital publishing environment. Because of this, their productions — personal web pages and online course packets — are difficult to maintain, are not readily interoperable or standards-compliant, and are easily dismissed as heterogeneous grab-bags of links. NINES was founded to work against this debilitating situation.

The inherent complexity of available resources is a further obstacle to the penetration digital humanities into the disciplines. *Collex* is designed to aid humanities scholars doing research in complex digital collections like the *Rossetti Archive* (its initial test case) or within federated research environments like NINES. Such environments often stymie their users through the sheer quantity of information made available to them in top-level tables of contents, sitemaps, and idiosyncratic search engines. Our tool operates under the assumption that the best paths through a complex digital resource are those forged by use and interpretation. A *Collex* approach works to assist scholars in recording, sharing, and building on the interpretive purposes to which they put their online teaching and research environments.

Collex uses semantic web principles and technologies to explore and develop the research potential of the digital scholarship aggregated in NINES. Two critical concepts embodied in a NINES environment shaped by the *Collex* application fall under the rubrics widely known as "faceted classification" and "folksonomy." Facets and folksonomies structure an approach to descriptive metadata. They generate an evolving interface between the fully-integrated peer-reviewed electronic resources

that constitute NINES and the user communities that re-imagine NINES content through interpretation, contextualization, and critical and creative re-fashioning.

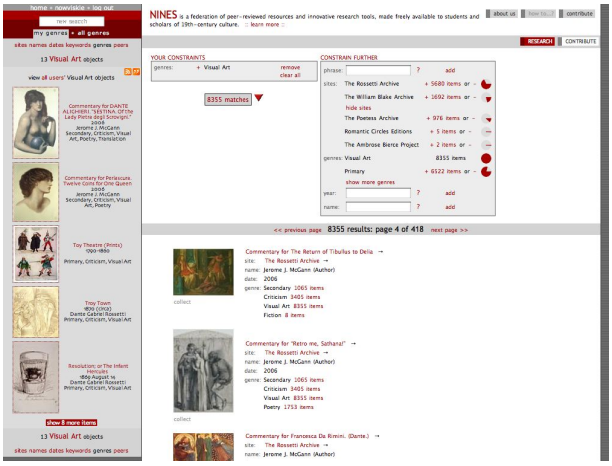
"Full integration" means that each of our NINES-participating resources has contributed a package of metadata representing all of the digital objects they wish to make browseable, collectible, and available to users for re-purposing within *Collex*. An important innovation of *Collex* lies in the way these objects are defined by their contributing editors. *Collex* uses a Dublin Core flavor of RDF, the "resource description framework" of the semantic web, to define collectible "objects" without limiting them to their expression as web pages. Where other social bookmarking tools are designed to allow collection and annotation of whole web pages, *Collex* allows contributors of resources to make finer-grained distinctions, and users of the system to build collections and exhibits more attuned to the patterns of attention in humanities scholarship.

A clear example of interpretive modeling through object definition is the *Collex*representation of a book of poetry in the Rossetti Archive. Using XSLT transformations, we have created RDF metadata for intellectual and material objects at differing levels in this book. One RDF object (typed as a secondary resource, with supplemental genre and date identifiers) expresses the editor's commentary on the book as a concept. Another object, also articulated in metadata, expresses one particular edition of the book. Within that high-level expression, each page of the book has been shared with *Collex* users as a collectible object, as has each poem on each page. Such fine disambiguation ensures that *Collex* users can locate, annotate, and exhibit objects specifically suited to the scholarship they wish to perform – whether their attention is focused on bibliographic, social, or textual matters. It also ensures that archive maintainers have the fullest control, in the *Collex* environment, over the use of their intellectual property and the artifacts they minister.

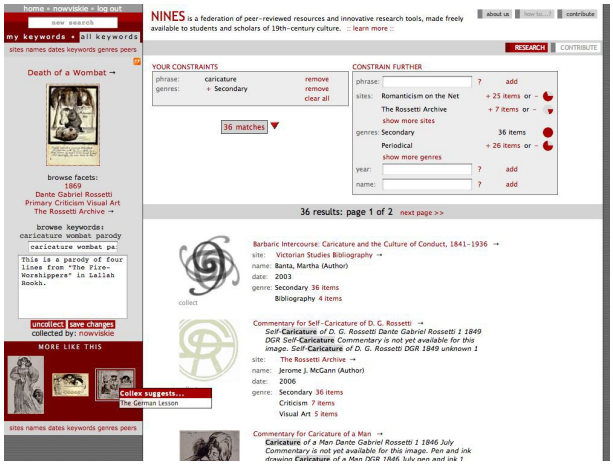
Because RDF objects share a common (and relatively simple) metadata scheme, they are discoverable through "facets" in the *Collex* search and faceted-browsing interface. Faceted classification is a non-hierarchical means of expressing ontological relationships. Any given object will share a number of facets with other objects – common dates, genres, authors, etc. Exposing these facets makes it possible not only for users to manually "drill down" into certain categories or explore lateral relationships, it also opens possibilities for algorithmic serendipity in research. In other words, *Collex* can exploit formally-expressed facets to offer more options and avenues to users interested in a particular object: "more like this" – more objects in the repository sharing one or more attributes with a researcher's subject of attention. Even more interesting is the ability of *Collex* to record and analyze user activity, and to translate the products of user interaction into RDF objects within the system itself. In this way, in addition to "more like this," *Collex* can suggest to recent collectors of a particular object that they view the published collections and exhibits into which other users have placed the object, or objects like it. Because this content can be expressed as subscription-based RSS feeds, a web service, or an API, it is possible for the maintainers of scholarly resources to patch into *Collex* directly from their individual web page or listserv interfaces, offering information about user annotations and re-mediations for any given object without requiring users to visit *Collex* at all.

All *Collex* activity takes place within the ordinary web-browsing environment that scholars presently use to access digital resources, and will require nothing in the way of plugins or downloads. The overhead (in terms of initial metadata production) for contributors of resources to the federated collections in which *Collex* can operate has also been kept purposely low, and is thoroughly compatible with Open Archives protocols. We predict that both of these factors – combined with the strong endorsement and example of NINES – will facilitate the adoption of *Collex* into day-to-day practices of humanities scholars in networked research and publishing environments.

Screenshots



A Collex sidebar list view (user-collected objects in the "visual art" genre) with the same constraint in the faceted browser



A different constraints set, and a detail view of one object (for tagging, annotation, and knowledge discovery) in the sidebar

Bibliography

"Access to the Literature: The Debate Continues." *Nature Web Focus* (2004). <http://www.nature.com/nature/focus/accessdebate/>

Broughton, V. "Faceted Classification as a Basis for Knowledge Organization in a Digital Environment; the Bliss Bibliographic Classification and the Creation of Multi-Dimensional Knowledge Structures." *New Review of Hypermedia and Multimedia* 7 (2001): 67-102.

Golder, Scott, and Bernardo Huberman. "The Structure of Collaborative Tagging Systems." *Information Dynamics Lab, HP Labs* (2005). <http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf>

Hammond, Tony. "Social Bookmarking Tools (I): A General Overview." *D-Lib Magazine* (April 2005). <http://www.dlib.org/dlib/april05/hammond/04hammond.html>

Lessig, Lawrence. *Free Culture.* . <http://www.free-culture.cc>

Lynch, Clifford A.. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." *ARL Bimonthly Report* 226 (2003).

Mathes, Adam. "Folksonomies - Cooperative Classification and Communication Through Shared Metadata." Paper written for LIS590CMC Computer Mediated Communication, Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, December 2004. 2004. <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>

McGann, Jerome. *Radiant Textuality: Literature After the World Wide Web*. Palgrave MacMillan, 2001.

McGann, Jerome. "Culture and Technology: The Way We Live Now, What is to Be Done?" Paper presented at the University of Chicago, April 23, 2004. 2004. <http://www.nines.org/about/bibliog/mcgann-chicago.pdf> from="ROOT" targOrder="U"/>

MLA. "The Future of Scholarly Publishing: Report of the Ad Hoc Committee on the Future of Scholarly Publishing." *Profession 2002* (2002): 172-186.

Nowviskie, Bethany. *Collex: Semantic Collections and Exhibits for the remixable Web* . <http://www.patacriticism.org/collex/about>

Nowviskie, Bethany, and Jerome McGann. *NINES: A Federated Model for Integrating Digital Scholarship* . White paper version available at <http://www.nines.org/about/9swhitepaper.pdf>

Unsworth, John. "Not-so-Modest Proposals: What Do We Want Our System of Scholarly Communication to Look Like in 2010?" Paper presented at the CIC Summit on Scholarly Communication, December 2, 2003. 2003. <http://www.iath.virginia.edu/~jmu2m/CICsummit.htm>

Unsworth, John. "Tool Time, or 'Haven't We Been Here Already?' Ten Years in Humanities Computing." Paper presented at Transforming Disciplines: The Humanities and Computer Science, NINCH conference, Washington, DC, January 18, 2003. 2003. <http://www.iath.virginia.edu/~jmu2m/carnegie-ninch.03.html>

Van de Sompel, H et. al. *D-Lib Magazine* (September 2004).

The Visionary Cross: An Experiment in the Multimedia Edition

Daniel Paul O'Donnell
(daniel.odonnell@uleth.ca)
University of Lethbridge
Catherine Karkov
University of Leeds
James Graham
University of Lethbridge
Wendy Osborn
University of Lethbridge
Roberto Rosselli Del Turco
Università degli studi di Torino

A reservation I should wish to express is that customarily levelled at digital projects, which is that while the technology (brilliantly, beautifully, wonderfully) enables and indeed encourages the presentation of multiple points of view, so putting the burden of interpretation onto the individual reader, there is a concomitant loss of genuine decision-making by those claiming to be responsible for the work as a whole... Editors should edit: will they?

Our epigraph comes from the report of an external assessor to the original funding application for our project. We cite it because it represents a common question asked of the Digital Humanities by traditional scholars: "Can it be as significant as it is pretty?"

For editors of text-based digital projects, the answer is increasingly clear. The last decade has seen the development of a relatively solid consensus as to basic technological and generic expectations (see O'Donnell 2004 for a summary). Best practice now expects that text-based digital projects will be encoded using XML, preferably TEI. It assumes they will contain an archive with transcriptions and full colour facsimiles of primary sources; that some means will be provided for comparing variant readings; and that users will be able to test editorial assumptions by comparing or constructing alternative editorial texts. While there is some debate as to whether text-based projects have yet lived up to their original promise (Robinson 2005), there can be little doubt they are beginning to be recognised as important works of scholarship in their own right.

For developers of projects that depend heavily on multimedia or collaborative technologies, however, the answer to this question is far less clear. While open standards exist for the encoding of image, moving pictures, and sound, there is little agreement as to how these are to be presented to the end user: unlike text-based projects, multi- and mixed media projects still commonly rely on proprietary software or specific operating systems (e.g. Foys 2002, Reed-Kline 2000; British Library Board, n.d.). And while many digital projects propose using collaborative technology in their design, there is as yet no agreement on the fundamental issue of how such collaboration can function in a research culture based on peer review and the preservation of authorial integrity. A number of exemplary projects are beginning to show how such technologies can be applied in specific contexts or to solve specific research problems (e.g. Ó Croinin et al. [n.d.], Toth et al. [n.d.]). But we are still far from agreeing as to how they can be used more generally to support day-to-day research by working humanities scholars.

The Visionary Cross project addresses this problem by treating it as a research question. Our goal is to produce a mixed-media and extensible edition of a key group of Anglo-Saxon artefacts associated with the "Visionary Cross" tradition in Anglo-Saxon England: the eighth-century Ruthwell and Bewcastle standing stone crosses, the tenth-century Vercelli Book dream of the Rood poem, and the eleventh century Brussels Reliquary Cross (for this tradition, see Ó Carragáin 2005).

These objects include some of the best known and most studied of the period. The Ruthwell Cross is a 17 foot high stone cross erected near a former Roman military site in Dumfriesshire Scotland. It is perhaps best known to Anglo-Saxonists for a runic inscription that may be the oldest known record of an Anglo-Saxon vernacular poem, versions of which can be found in the tenth-century Vercelli Book and eleventh-century Brussels Cross (see Ó Carragáin 2005, 58-60; O'Donnell 1996, 287-288, for bibliography).

The Brussels Cross is a reliquary that once contained a fragment from the supposed True Cross. It is built on an oak core that was covered with precious metal and jewels and perhaps a crucifixion (stolen sometime before 1793; see Van Ypersele de Strihou 2000; Webster 1984; Ó Carragáin 2005). Gilt silver decoration on the cross's back and side bands bearing a vernacular inscription have survived. On the centre of the back of the cross is a depiction of the Agnus Dei; symbols of the four evangelists are found at the terminals. An Old English inscription around the edge quotes from a version of the same poem found on the Ruthwell Cross and in the Vercelli Book. A second inscription explains that the cross was made by two

brothers in memory of a third. On the back we are told the name of the artist responsible for its manufacture.

The Bewcastle Cross is a standing stone cross found, like the Ruthwell Cross, at a former Roman military site. Approximately the same size as Ruthwell and belonging perhaps to the same artistic school, the severely weathered Bewcastle Cross still stands in its original location (see Bailey and Cramp 1988). It has the remains of a sundial on its side and may have been painted and decorated with other metalwork or glass attachments. The west face is carved with three figural panels, of which two also appear on Ruthwell. The east side of the cross is decorated with a continuous vinescroll similar to Ruthwell; its north and south sides are carved with panels of interlace, geometric, and foliate ornament. The lowest panel on the west face shows a falconer wearing secular dress. This usually is understood to represent the deceased man commemorated in a now largely illegible runic inscription.

The Vercelli Book *Dream of the Rood* poem ties the members of this collection together (ed. Swanton 1996). The *Dream* poem describes an encounter with an object that is at once and alternately a tree, a beacon (a word used to describe the Cross on the Bewcastle Cross), a sign, and a cross sometimes covered with blood, and sometimes covered (as in Brussels) with gold and jewels. It ends with the Cross instructing the dreamer to tell what he or she has seen and with the dreamer reciting an expression of devotion and commemoration. The *Dream* is one of only about 25 poems and poetic fragments known to have survived the Anglo-Saxon period in more than one copy (O'Donnell 1996; see also Orton 2000). If the runic carving on Ruthwell is coeval with the rest of the monument, then the poem has a textual history that is longer and more geographically and linguistically diverse than almost any vernacular poem in the period. The citation of a couplet from the text on Brussels, moreover, suggests that it occupied a very significant place in the vernacular literary imagination: the only other known example of a similar verse citation in the period is from the translation of the Psalter.

Together, these objects form a cultural matrix whose members are associated along a number of textual, art historical, liturgical, and archaeological planes. The goal of this project is to use new technology to study these objects and their relationships in ways impossible in print—or even in person. Just as a textual edition improves upon witnesses by contributing an interpretive apparatus, so to our edition will improve on readers' knowledge of this matrix by placing it in a hypermedia apparatus that will assist in its interpretation.

The value of this approach is perhaps most obvious in the case of the crosses, which can be understood as multimedia objects in their own right. In all three cases, the monuments gain meaning from the interaction of text, image, and context. The stone crosses appear to have been “read” by walking around in a direction determined by their geographical orientation and the order of the Liturgy. The Brussels cross—depending on one's view of the object's original function—would likely have been seen by contemporary audiences either as an altar piece or carried in procession (On this spatial aspect see especially Ó Carragáin 2005).

The new technologies also allow us to ask new questions about the objects relationships with each other. Had an Anglo-Saxon observer been lucky enough to see all four in a single lifetime (an impossible proposition given their temporal and geographic distribution), he or she would have understood them both as individual works of art and as part of a larger web of cultural traditions and references extending along various textual, art historical, and generic planes. By taking advantage of hypermedia's strength in the representation of arbitrary connections, we as editors can now represent these connections to modern scholars in a way that translates and augments the original artefacts—in our edition, linking becomes a type of hypermedia collation. In our edition, scholars will be able to both to study the individual objects as objects in their own right and follow the connections among them. In doing so they will have access both to a collection unavailable to any single Anglo-Saxon observer and the benefit of immediate access to the best of recent criticism and centuries of secondary scholarship.

By using recent developments in collaborative technologies, finally, we hope this project—like the cultural knowledge it attempts to capture and represents—will be open to augmentation as our knowledge develops. By using standoff markup, we intend to allow developers and users to anticipate connections to other objects in the matrix or discover new connections among existing objects in much the same way contributors to the Wikipedia can predict the existence of articles that have yet to be written or contribute “stubs” for subsequent elaboration while retaining intellectual ownership of their contributions (see Ore 2004 for a discussion of collaborative editing; O'Donnell 2006 discusses some strengths and weaknesses of the model for scholars).

If multimedia projects are going to answer our reviewer's question, they must learn to do more than simply display—they must also learn to *edit*. This paper discusses the approaches we are and will be taking to this important problem in developing a complex multimedia “edition” of a cultural matrix.

Bibliography

British Library. *Turning the Pages*. n.d.. <http://www.bl.uk/collections/treasures/digitisation4.html>

Foys, Martin K., ed. *The Bayeux Tapestry.* Leicester: Scholarly Digital Editions, 2003.

Ó Carragáin, Éamonn. *Ritual and the Rood: Liturgical Images and the Old English Poems of the Dream of the Rood Tradition.* Toronto: University of Toronto Press, 2005.

Ó Cróinín, Dáibhí, et al. "Profilometry of Medieval Irish Stone Monuments." *Foundations of Irish culture AD 600-850.* n.d.. <http://www.foundationsirishculture.ie/main.php?id=4>

Ore, Espen E. "Monkey Business - or What is an Edition?" *Literary & Linguistic Computing* 19.1 (2004).

O'Donnell, Daniel Paul. "Manuscript Variation in Multiple Recension Old English Poetic Texts: The Technical Problem and Poetical Art." PhD Thesis. Yale University, 1996.

O'Donnell, Daniel Paul. "The Doomsday Machine, or, 'If You Build it, Will They Still Come Ten Years from Now?'." *Heroic Age* 7 (2004). <http://www.heroicage.org/issues/7/ecolumn.html>

O'Donnell, Daniel Paul. "O Captain! My Captain! Using Technology to Guide Readers through an Electronic Edition." *Heroic Age* 8 (2005b). <http://www.heroicage.org/issues/8/em.html>

O'Donnell, Daniel Paul. "Why Should I Write for your Wiki?" Working Paper 1. Readex Community Academic Advisor Board. 2006.

Reed Kline, Naomi, ed. *A Wheel of Memory: The Hereford Mappamundi.* Ann Arbor, MI: University of Michigan Press, 2001.

Robinson, Peter. "Current Issues in Making Digital Editions of Medieval Texts—or, Do Electronic Scholarly Editions Have a Future?" *Digital Medievalist* 1.1 (2005). <http://www.digitalmedievalist.org/article.cfm?RecID=6>

Swanton, Michael James. "The Dream of the Rood." *Old and Middle English Texts.* Exeter: University of Exeter, 1996.

Toth, Michael B., et al. *The Archimedes Palimpsest.* n.d.. <http://www.archimedespalimpsest.org/>

Van Ypersele de Strihou, Anne. *Le Trésor de la Cathédrale des Saints Michel et Gudule à Bruxelles.* Brussels: Cathedral, 2000.

Webster, Leslie. "The Brussels Cross." *The Golden Age of Anglo-Saxon Art, 966-1066.* Ed. Janet Backhouse, D. H. Turner and Leslie Webster. London: British Museum, 1984.

The LInguistic and Cultural Heritage Electronic Network (LICHEN): A New Electronic Framework for the Collection, Management, Online Display, and Exploitation of Multimodal Corpora

Lisa Lena Opas-Hänninen *(lisa.lena.opas-hanninen@oulu.fi)*
University of Oulu
Finland

Matti Hosio
Univerisity of Oulu
Finland

Ilkka Juuso
University of Oulu
Finland

Tapio Seppänen *(tapio@ee.oulu.fl)*
University of Oulu
Finland

The international, interdisciplinary and multilingual LICHEN project, initiated by the Department of English and the MediaTeam research group (Dept. of Electrical and Information Engineering, MediaTeam 2006) at the University of Oulu and the SCOTS corpus project at the University of Glasgow (Scottish Corpus of Texts and Speech, 2006), focuses on the languages and cultures of the northern circumpolar region. Its underlying assumption is that language and culture are as important to the survival and well-being of populations as more obvious ecological, social and health issues and thus it is also a member of the Circumpolar Health and Wellbeing research programme run by the Centre for Arctic Medicine, University of Oulu (Thule Institute 2006).

The aim of the project is two-fold: firstly, the project aims to collect, preserve and disseminate information about the languages spoken in the circumpolar region, thus also enabling research on them. This will also help to promote the linguistic

confidence and self-image of the speakers of these languages, strengthening their cultural awareness and facilitating cross-cultural communication between these peoples in an age of rapid global change (Winsa 1998).

Secondly, and more importantly, the project aims to create an electronic framework for the collection, management, online display, and exploitation of existing corpora of the languages of the circumpolar regions, which is also applicable to other corpora that represent regional, social and other varieties of languages. Humanities computing researchers, in particular, have long recognized the need for new, more sophisticated tools to aid scholarly research of textual data, not to mention tools that would be able to handle multimodal data. Although a number of tools have been developed, they suffer from various restrictions, e.g. they are only applicable to the data they were developed for, importing data is laborious, user interfaces and encoding standards are outdated, considerable expertise in programming is assumed, no support for multilinguality is included, or they promise more than they offer. While there have been some very promising advances made in this direction (e.g. TAPOR Tools 2006), it is clear that more tools are needed.

The framework being developed in this project is intended to be the equivalent of an extendable toolbox for corpus linguists. It will attempt to offer much-needed functionality in an easy-to-use package, which is shaped and built-on according to real user needs. Initially emphasis will be given to the implementation of the text capabilities of the system, but other modalities (such as audio and video) are also taken into account. The idea is to facilitate queries into a multimodal database using both proven and novel ways of finding and displaying information (Seppänen 2006). Metadata and metadata visualisation, particularly in conjunction with the new modalities, will be essential in achieving this. While we support the use of best practices for the collection, preservation and presentation of corpus data, we also recognize that some data, particularly legacy data, may not be in a position to do so and the shell must also support such data (Kretzschmar et al. 2006).

The system will also make migration to and from other tools straightforward by offering import and export features for commonly used programs. It will enable users to bring in their own data, which they can keep private or make public using the built-in web functionality. The database will also be capable of handling several different versions of any document (for example, revisions, interpretations or translations); these are linked, a feature that can be made use of in queries. Queries can be made using regular expressions, which may combine free-form text (words, phrases) and part-of-speech tags, for example.

The system is implemented in Java making it platform independent and taking advantage of the many technological components developed for that language. The ultimate goal of the development of the computing tools is a shell which can be adapted to any language. Therefore, support for multiple languages and a variety of character encoding schemes are important.

The main focus of the project is on Meänkieli and Kven, two Finnic minority languages spoken in Sweden and Norway, respectively, and Scots and Scottish English. At present we have about 150 hours of tapes in Meänkieli and 100 hours of tapes in Kven. More Kven material is currently being collected. We also have access to both the structure and contents of the Scottish Corpus of Texts and Speech (SCOTS) at the University of Glasgow, currently totalling 3.5 million words of spoken and written Scots and Scottish English.

The project began in 2004 and a prototype of the shell has now been constructed. We will demonstrate this shell, showing some of the basic functionality of the system while looking into concrete research questions focusing on the image of Scottishness as presented in Irvine Welsh's *Trainspotting* (1996). Since the story is found in novel, play and movie versions, it affords an excellent testbed for a toolbox that can handle multimodal data. Using a few key scenes as examples, our research focuses on the questioning of national stereotypes in terms of landscape, language and culture. We are thus interested in the comparison of the images presented in the three versions, which will also demonstrate the ability of the toolbox to support versioning. While we concentrate on linguistic features of Scottish English, we also demonstrate how easy-to-use access to sound and images linked to the transcription of the movie and the linking between the three versions of the text greatly facilitate research such as ours that must take into consideration images created both on all levels of a text and across texts. Finally, we demonstrate the possibility of making use of online dictionaries as an added tool in the analysis of data from within the toolbox.

Since the shell is still in prototype form, we would welcome this opportunity to discuss our needs and goals at DH2007, thus drawing on the considerable expertise of the conference participants in order to ensure that our tools benefit as wide a range of users as possible.

Bibliography

TAPOR Tools . . Accessed 2006. <http://tapor.humanities.mcmaster.ca/home.html>

Trainspotting. Dir. Danny Boyle.

4 Play. Vintage, 2001.

Hodge, John. *Trainspotting.* DVD. Vintage, 1996.

Kretzschmar, W. A. Jr., J. Anderson, J. Beal, K. Corrigan, L. Opas-Hänninen, and B. Plichta. "Collaboration on Corpora for Regional and Social Analysis." *Journal of English Linguistics* 34.3 (2006): 175-205.

MediaTeam. *Oulu Research Group* . . Accessed 2006. < http://www.mediateam.oulu.fi/brief/?lang=en>

Scottish Corpus of Texts and Speech . . Accessed 2006. <http://www.scottishcorpus.ac.uk/>

Seppänen, Tapio. "Multimedia Information Retrieval." Plenary talk at Digital Humanities 2006, Paris, 5-8 July 2006. 2006.

Thule Institute . . Accessed 2006. < http://thule.oulu.fi/>

Winsa, B. "Language Attitudes and Social Identity. Oppression and Revival of a Minority Language in Sweden." *Applied Linguistics Association of Australia* 17 (1998).

The Role of the Computer in Humanities Computing

Wilhelm Ott *(wilhelm.ott@uni-tuebingen.de)*

Tübingen University

Professor Ott delivered the 2007 Busa Prize Lecture under this title. What follows is the introduction to the lecture, provided by Michael Sperberg-McQueen.

As John Unsworth has just said, the fourth Busa Award is given to Wilhelm Ott. The members of the committee for this cycle of the Busa Award were:

- Lorna Hughes (chair)
- Lisa Lena Opas-Hänninen
- Espen Ore
- Steve Ramsay
- Stefan Sinclair

The members of the Award committee have asked me to say a few words about the work of Wilhelm Ott.

Like many of us, Wilhelm Ott came to our field by a circuitous route.

As a young man, he studied at the Collegium Germanicum, a seminary in Rome, but, if I understand correctly, he experienced some tensions with his superiors. As in the traditional university, a four-year study of the liberal arts was required before students were allowed to embark on the study of the queen of sciences, theology. Herr Ott and some of his fellow students incurred the displeasure of the authorities by sneaking across the road in some otherwise unoccupied time, to sit in on a series of lectures on Paul's epistle to the Romans.

And then, in his master's thesis, that chapter on Heidegger! Well, really.

Let's say just that when he continued his education elsewhere, it was a mutual decision.

It was natural, perhaps, that from theology he should turn to classical philology. As a student of modern languages, I have always envied classicists their extensive set of ancillary tools, their lexica, their dictionaries, their handbooks.

But to someone coming to classics from theology and Biblical literature, of course, what seems most striking classics is the striking poverty of the instrumentarium. Specialized glossaries

and grammars for individual authors, concordances, systematic lists of all kinds were much less thick on the ground than one might wish.

And so, characteristically, he set out to help fill this gap.

His first work in this vein is a series of books published beginning in 1970 providing tables and data for the systematic study of the Latin hexameter: first of all a volume of Metrische Analysen zur Ars poetica des Horaz, which provides tables of scansions, elision, ictus, and related information for each line of Horace's poem, modeled on similar tables for metrical studies provided by Eduard Norden in his edition of the Aeneid.

Similar volumes followed, over the years, for the twelve books of Vergi's Aeneid, for the Georgica and the Bucolica, for Lucretius, and so on.

In the Foreword to the volume on Book VI of the Aeneid, there appears a sentence which seems to sum up, programmatically, both the capabilities and the potential limits of digital processing:

Electronic data processing can be put into service whenever data of any kind — notably including texts — must be processed according to rules which are unambiguously formulatable and completely formalizable.
What more compact formulation could we find of the fundamental program of our field? And what more matter of fact reminder that this is a description of those places where computers can successfully be deployed, without any suggestion of belief that they can be deployed absolutely everywhere.

The metrical tools set an emphasis which has continued to characterize all of Herr Ott's work throughout his career: computers should serve scholarly purposes, not vice versa. In his discussion of the tables and their preparation, three virtues are seen as particularly important:

- completeness
- reliability
- verifiability

All of these concerns characterize Herr Ott's later work as well.

There is, too, a concern for accessibility. Initially, it motivated the publication of the metrical analyses in book form, to make them accessible to classicists without compuational expertise or means. Later, the concern for accessibility led to efforts to make the software which generated those analyses accessible to other scholars, to extend them to make a more tool useful for other kinds of analysis as well. In due course, these efforts produced a suite of programs for scholarly work with text, which at some point acquired the name Tustep, the Tübingen System of Text Processing tools.

Tustep embodied a number of important ideas:

- completeness: It can be used for all parts of a project's normal work. There is an editor for data capture and revision, there are copy commands and tape utilities for archiving and moving data, there are a variety of general and specialized proceessing tools for manipulating documents, for sorting things, for extracting relevant items from lists, for laying documents out on the page, for photocomposition of the resulting pages, and so on.)
- verifiability: Every action undertaken with Tustep will be logged, unless you take very active steps to avoid having it logged.
- stability: Since projects may live for years or decades, the stability of the program and of Tustep files is critically important.
- consistency: Years before anyone outside of Bell Labs had heard of Unix, Tustep adopted the principle that every tool would have one primary input and one primary output, and that the output of every tool would be usable as the input to any other tool. In practice, this means that the primary input and the primary output of each tool use the Tustep file format.

This principle of ensuring that other Tustep programs can read the primary output of any Tustep program is consistently implemented, even in cases where one might have expected a different choice. When I learned that the typesetting program of Tustep also produces a Tustep file as its output — the PDF or photocomposer file is, formally speaking, a side effect — I admired the consistent application of the design principle but privately thought that it bordered on the academic. Since for practical purposes the main output of a typesetting program is typeset pages, producing a Tustep file seems likely to be an anticlimax. What useful output can it produce? Perhaps just a copy of its input?

The most important idea of Tustep, though, is that it is the responsibility of the software to serve the needs of scholarship, and not vice versa, and that the responsibility of the scholar is to respect the significant particularities of the material and the demands of his discipline (not any standards of practice imposed from outside, and least of all any limitations imposed by the software.)

Tustep developed over thirty years of listening to the needs of scholarship, consulting with projects and adding to Tustep the functionality they needed to enable them to do their work. Hundreds of editions have been prepared with it, some all the way from beginning to end, from data capture through typeset pages, others just translated into Tustep for the typesetting — apparatus criticus is not easy to set!

If we are to take responsibility, as humanists, for our use of machines, then it is necessarily now a part of humanities scholarship to understand and develop ways to make machines

adapt to the requirements of our work, and (while remaining open to the exploitation of new and unforeseen opportunities) to resist the temptation to adjust our practices to suit the convenience of the machine. In this sense, Wilhelm Ott's decades of work on Tustep have been not only the work of a software developer, but more profoundly the work of a gifted humanist.

It has been the great good fortune of our field to benefit from Wilhelm Ott's work as a scholar. His work has taught a great deal over the years to those wise enough to learn from it. And I for one am grateful for the chance, this evening, to learn from him in person.

Bringing the Digital Revolution to Judaic Music: The Judaica Sound Archives (JSA)

Salwa Ismail Patel *(spatel37@fau.edu)*
Florida Atlantic University

The Judaica Sound Archives at Florida Atlantic University (FAU) has been in existence since 1990s with primary funding from the FAU Libraries and some private funding from donors. It was started as a grassroots effort by a Cantor to just preserve and save some cantorial music and has now grown to encompass all the arrays of Judaica music including as many as 70,000 recordings. JSA's aims and objectives have evolved from the initial goal of establishing a preservation-quality digital archive for at-risk sources of Judaic music to a full online delivery system for sound, images and metadata with tools for scholarly research and annotation. However, what has not wavered or changed is the total commitment to the highest quality and resolution for sound and images, excellent work practices and rigorous quality control and constant evolution and adaptation to new evolving technologies for audio and textual digitization.

This abstract describes the JSA project from its inception to its current phase, showcasing the innovative work that is being done to develop this project, using computer technology to overcome certain challenges and the usage of computing in a unique method to make this collection more accessible to a wider range of scholars and students.

JSA came about as a result of a fortuitous collision between of a corpus of music recordings being donated to the FAU Libraries and the rapid advances in digital technologies at FAU Libraries making music accessibility to the academic users a possibility. All these recordings are primary sources that revolve around the life related to Jewish and Yiddish/Hebrew speaking cultures in the United States in the early and mid-years of the 20th century. The principal work of this project falls into two broad categories: digitizing rare at-risk recordings and making these available digitally to the scholars, researchers and students interested in computing humanities sound recordings. Beginning with less than a thousand 78 rpm recordings, JSA's holdings have expanded to include LP albums, cassette tapes and 45 rpm records. Taken together, this represents well over 70,000 sound tracks. The collection is continually growing as a result of material donations from institutions and individuals. The 78 rpm sound recordings received at JSA are identified by song,

performer, and composer and entered into a database which is searchable on this website. At present the collection holds 3,450 different titles on 78 rpm records, representing almost 7,000 different sound tracks produced between 1901 and the mid-1950s. And JSA's current short term goal is to digitize the best example of every recording in the collection. The original phonograph records or tapes are saved, shelved and protected from further damage. One digitized copy is filtered and enhanced electronically to improve the quality of the sound. Another is kept in its original form.

According to, Samuel Brylawski, head of Recorded Sound Section at the Library of Congress," Digital media have the advantage of not suffering any loss of information as they are copied, unlike the generational losses inherent in the duplication of analog media such as discs and cassette tape. The future of audio preservation is reformatting audio tapes and discs to computer files and systematically managing those files." University of California reported the first digital audio project in the 1980s. Since then there has been a steady increase in digital audio preservation projects. In a survey conducted by Richard Griscom, about 42 libraries in the US responded to having some kind of digital audio project available. However, JSA is the only project in the nation that works on not only preserving, but also digitizing and making accessible online and on-site the rare Judaic recordings. Other important work in the field of audio digitization is being done by Cornell University, Indiana University and University of California, Berkeley. However, Indiana University is the only library system that is dealing with audio that is music much like JSA. The difference in their approach is that it is not specific to a certain culture or region and they are more focused on music images and sheet music scores unlike the JSA focus on making music files accessible and available. The music is searchable using a database that implements the Daitch-Mokotoff Soundex encoding. The music is then streamed onto the listeners' computer.

This poster discussion will also include the various steps involved in digitizing audio. We will discuss the Analog Format Inspection, Playback Equipment Calibration, Analog to Digital Conversion, Digital Editing System, and Accessibility and Distribution. For the digital media, the goal is to make the collection as generic as possible, thereby maximizing accessibility and setting the stage for easy migration to the next generation of digital storage.

In conclusion, this poster will provides a detailed overview of the critical steps relevant to the digitization process of these analog primary resource materials and making this music available and searchable online.

The Encoding of Time in Manuscripts Transcription: Toward Genetic Digital Editions

Elena Pierazzo *(elena.pierazzo@kcl.ac.uk)*
King's College London

Writing is a process that occurs in time. This simple and obvious consideration involves many issues both from a theoretical and practical point of view. A critical evaluation of timing is really crucial in the case of modern authors' autograph draft manuscripts because different layers of corrections, deletions and additions can give insight into an author's way of working, key to the interpretations of his/her works and the evolution of the author's *Weltanschaung*, as highlighted in genetic criticism over the last few years.[1] Medieval manuscripts copied by one ore more scribes are also, of course, the result of a process that occurred in time, but the different kind of authorship involved in such cases seems to involve a difference in the evaluation of the cultural weight of recorded variants.

When a scholar inspects a written text, especially a manuscript, s/he has in his/her hands the final result of that process and can choose whether to approach it from a codicological/documentary point of view, "photographing" the resulting product, or from a genetic point of view, trying to describe the teleological flow of authoring. The first approach is more typical of scribes' copies, the latter of autographic (draft) manuscripts.

Practically, time based editions are really difficult to represent in printed works, because of the bi-dimensionality of paper sheets.

> On voudrait représenter dans la bidimensionalité des pages un processus génétique dont on s'est pourtant appliqué à montrer que sa proprieté est d'ajouter à l'écrit, qui est bidimensionel, une troisième dimension, qui est celle du temps!
>
> (Gresillon 1994:121)

In a digital framework bi-dimensionality can be overruled by a hypertextual/multimedia approach that can allow the creation of a more flexible context for the presentation of a genetic edition. The problem of a time based encoding has been discussed in several circumstances (i.e. TEI Manuscripts SIG Meeting Report 01, Vanhoutte 2002 that suggests the usage of the markup solution employed in the transcription of speech for the purpose), but – up till now – a coherent encoding model

for such editions has not been proposed. And for a good reasons: while it is possible to describe a relative-timed process, it is very complicated, if not impossible, to draw a general absolute-timed framework.

Analysis and time-based encoding of authorial interventions

Let us consider a couple of examples taken from page 3595 of Zibaldone of Giacomo Leopardi.

Fig. 1

Transcription: che e' si rechi a' ~~dotti~~ denti l'un d'essi cibi

In this line we can detect two corrections: 1) deletion of the line under “si” and 2) deletion of “dotti” corrected into “denti”.

We can try to draw the timing of the creation of the segment as follow:

Time 1: writing “che e’ si rechi a dotti”
Time 2: deletion of “dotti”; consequent writing of “denti”
Time 3: underlying of “si”
Time 4: deletion of the line under “si”
Time 5: underlying of “rechi” and “denti”
Time 6: writing of “l’un d’essi cibi”

Other timetables are also possible, but let’s assume this is the more probable one. The text can be encoded using a TEI-based mark-up with the help of a new global attribute (`@time`), intending that it’s just an example to help the conceptualization of the problem:

```
<seg time="1">che e' <del
type="underline deletion" time="4">
<hi rend="underline" time="3">si</hi>
> </del>
<hi rend="underline" time="5">rechi a'</hi>
<del type="overstrike" time="2">dotti</del>
</seg>
> <seg time="2">
<hi rend="underline" time="5">denti</hi>
</seg>
<seg time="6">l'un d'essi cibi</seg>
```

Such transcription tries to model the real flow of writing, but such a model may not be workable. In fact, it will fragment the flow of the plain writing in potentially infinite pieces. To simplify it, we can assume as Time 0 (default) the time of the normal plain writing flow, timing just editorial interventions.

The schedule will then be modified as follows:

Time 0: writing “che e’ si rechi a dotti”
Time 1: deletion of : “dotti”
Time 0: writing of “denti”
Time 2: underlying of “si”
Time 3: deletion of the underline under “si”
Time 4: underlying of “rechi” and “denti”
Time 0: writing of “l’un d’essi cibi”

A further simplification is also possible: assuming that – in genetic criticism terms – so-called “writing variant” (deletion of a single word substituted by another that immediately follows on the same line) occurs during the normal writing flow, the following model can be drawn:

Time 0: writing “che si rechi a dotti”; deletion of : “dotti”; writing of “denti”; writing of “l’un d’essi cibi”
Time 1: underlying of “si”
Time 2: deletion of the underline under “si”
Time 3: underlying of “rechi” and “denti”

This will be the consequent new encoding:

```
che e' <del type="underline deletion" time="2"> <hi
rend="underline"  time="1">si</hi></del>
<hi rend="underline" time="3">rechi a'</hi>
<del type="overstrike">dotti</del> <hi
rend="underline" time="3">denti</hi> l'un
d'essi cibi
```

The last possibility should not imply that any inline correction is to be considered as done in Time 0, but just the one followed by the correction. In fact, in the case of a deletion of an adjective or of any other word not essential from a syntactical point of view, the correction can occur in any time.

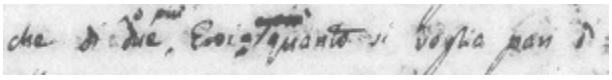


Fig. 2

Transcription: che di due °ro Eroi a °ro quanto si voglia pari di

This passage can be timed as follow:

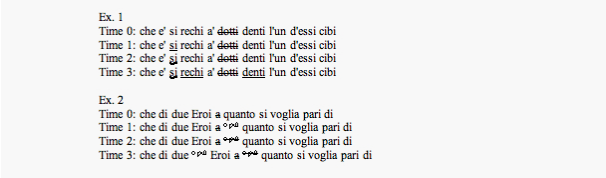
Time 1: writing “che di due Eroi a”;
Time 2: deletion of “a”;
Time 3: writing of “quanto si voglia pari di”
Time 4: interlinear addition of “o più” after “Eroi”
Time 5: deletion of the addition
Time 6: interlinear addition of “o più” after “due”

Or, in the simplified version, as follow:

Time 0: writing of “che di due Eroi a”; deletion of “a”; writing of “quanto si voglia pari di”
Time 1: interlinear addition of “o più” after “Eroi”
Time 2: deletion of the addition
Time 3: interlinear addition of “o più” after “due”

The encoding model (simplified version): `che di due <add place="intralinear" time="3">o più</add> Eroi <del type="overstrike">a</del> <del type="overstrike" time="2"><add place="intralinear" time="1">o più</add></del> quanto si voglia pari di`

By applying different stylesheets to the encoded texts, it will be possible to show all the different stages and to give the user the possibility of browsing among them.


Ex. 1
Time 0: che e' si rechi a' ~~dotti~~ denti l'un d'essi cibi
Time 1: che e' si rechi a' ~~dotti~~ denti l'un d'essi cibi
Time 2: che e' si rechi a' ~~dotti~~ denti l'un d'essi cibi
Time 3: che e' si rechi a' ~~dotti~~ denti l'un d'essi cibi

Ex. 2
Time 0: che di due Eroi a quanto si voglia pari di
Time 1: che di due Eroi a ~~°ºª~~ quanto si voglia pari di
Time 2: che di due Eroi a ~~°ºª~~ quanto si voglia pari di
Time 3: che di due °ºª Eroi a ~~°ºª~~ quanto si voglia pari di

Relative or absolute?

The two above examples occur on the same page: shall we then consider Time 1 of the first example the same of Time 1 of the second example? The answer should be: no; very little can be said about the timing of editorial/authorial intervention in two different segments. The possibility of establishing an absolute timing for correction is applicable only where we have strong palaeographic evidences or authorial declarations dating or describing a revision.

For instance, we can imagine an author used to typewrite his/her texts and then to correct them by hand: in this case the assumption of an absolute time is possible. But as different layers of hand corrections can also occur, there will be, in this case also, the necessity of considering relative-timed interventions. This situation can be represented in encoding distinguishing absolute and relative timing for instance by the application of two different attributes, i.e. `@timeRel` and `@timeAbs` In a digital framework an incorrect consideration of time as absolute or relative can bring to display texts that never existed. Let us imagine for a moment that we build an XSLT based tool able to display at a time either Time 1, Time 2 or Time 3 etc. variants: the results would be the display of variants that might have lived in different moments.

I think that, in absence of explicit authorial declarations or of palaeographic evidence, the only possible display would be to show timing of variants segment by segment, i.e. to give evidence just of relative timed corrections.

In the presentation I will present some examples from different authors (Giacomo Leopardi, Jane Austen, Virginia Woolf and some others) to examine the complexity of representation of temporal factors in digital genetic editions.

1. The question is too complex just to try to give some basic references; anyway Gresillon 1994 will offer a good starting point.

Bibliography

Grésillon, A. *Eléments de critique génétique. Lire les manuscrits modernes.* Paris: Presses Universitaires de France, 1994.

Sperberg-McQueen, C. M., and Lou Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange, XML Compatible Edition.* Oxford: TEI Consortium, 2002. `<http://www.tei-c.org/P4X/index.html>`

TEI Manuscripts SIG. *Meeting Report 01*. Nancy: ATILF, November 8th 2003 . `<at http://www.tei-c.org.uk/Activities/SIG/Manuscript/mssigr01.xml.ID=body.1_div.3>`

Vanhoutte, Edward. "Putting Time Back in Manuscripts. Text Ontology, Critique Génétique and Modern Manuscript." Paper presented at ALLC/ACH 2002 Tübingen: University of Tübingen, 25 July 2002. 2002. `<http://www.uni-tuebingen.de/cgi-bin/abs/abs?propid=93>`

ACH Panel: Employment - Pedagogy - Professionalization

Wendell Piez *(wapiez@mulberrytech.com)*
Mulberry Technologies, Inc.
Stephen Ramsay *(sramsay@unlserve.unl.edu)*
University of Nebraska - Lincoln
Geoffrey Rockwell *(georock@mcmaster.ca)*
McMaster University
John Unsworth *(unsworth@uiuc.edu)*
University of Illinois at Urbana-Champaign
Katherine L. Walter *(kwalter1@unl.edu)*
University of Nebraska - Lincoln

What sort of jobs are there in the field of Digital Humanities and where are they? How do you prepare for and get one? When you get one, what do you do? If you teach, what and how? How can you make sure your work is properly assessed? Where do you fit within the disciplinary world of the university? Panelists from industry, research centers, and academic programs will discuss these and related issues briefly, leaving plenty of time for a general discussion.

Form and Format: Towards a Semiotics of Digital Text Encoding

Wendell Piez *(wapiez@mulberrytech.com)*
Mulberry Technologies, Inc

Theories of the sign

This consideration begins in the theories of signification proposed in Structuralism, particularly as elucidated by Ferdinand de Saussure and interpreted by Roland Barthes (drawing on Hjelmslev and others). It will further be informed by the general theory of "autopoiesis" as articulated by Maturana and Varela (1987) and by other recent studies of language and signification influenced by systems theory.

Of particular importance to this treatment are the following Structuralist principles:

- Composite nature of the sign - a *sign* is an event or a relation between two components, a signifying part (*signifier*, *expression*, or *token*) and a signified part (the *content* or *meaning* of the sign).
- Arbitrary nature of the sign - the relation between signifier and signified in a sign is arbitrary, not inevitable or given by nature. Thus slippage is possible in principle and, in some systems, is common. This potential for slippage is what accounts for much of the flexibility, adaptability and power of sign systems.
- Signs work in combination: individual words, for example, have their significations, but when words are combined into sentences they become more useful and more expressive.
- Complete signs (considered as signifier/signified pairs) can also enter into signifying relations. For example, a *metalanguage* is a system of signs that describes another system of signs; expressions in the metalanguage represent signs in the signified sign system. Metalanguages provide channels of regulation and feedback that are conducive to the development of the sign systems they describe, and eventually (when formalized sufficiently) enable automated processes to manipulate signs systematically. Conversely, a *connotative system* is a system of signs in which the signifying parts of signs are themselves signs (one might think of literary texts). Connotative systems demonstrate the reciprocal relations between layers in a sign system and

the way significations at higher levels can condition and affect signification at lower levels, even apart from the application of metalanguages. While metalanguages abstract and formalize the sign systems they describe, connotative systems work by deploying signs as they are used concretely in other contexts, bringing alternative significations into play. Likewise, while metalanguages systematize and formalize, and thus indicate where automated rules-based processing is possible, connotative systems draw on and reflect particular significations available only in specific local contexts, and indicate where automation by traditional means is difficult or impossible.

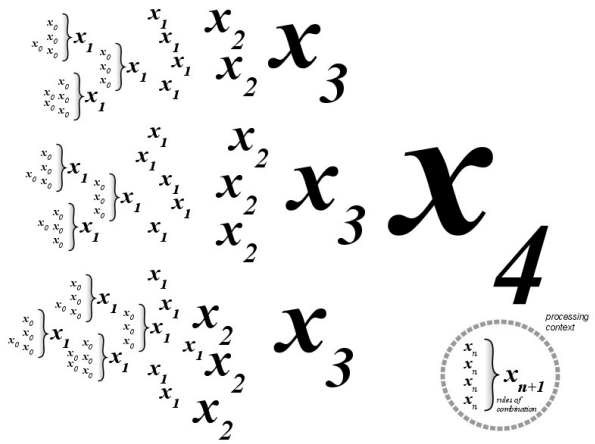


Figure 1: In both natural and artificial systems, a combination of simplicity of design with complexity, versatility and adaptability of application is achieved through a layered structure in which components with distinctive functions are made by combining simpler components.

The Semantics of Layers and the Layering of Semantics

A very useful distinction can be made between two different modes of signification, which we can call *operational* and *representational* semantics. The representational semantics of a sign corresponds to the conventional notion of what a sign is and how it functions, that a sign "stands for" something, reflecting and naming an actually existent "thing" in a real or imagined world. In contrast, the operational semantics describes not what the sign refers to in a purported world (which may or may not be present to the senses), but rather simply how it operates within the signifying system — which generally includes the world, or at least the present and active circumstances of the sign's use. (This in itself is the major difference between "human" and "machine" semantics [Piez 2002].) The rules of combination that allow any given set of signs to be assembled into a higher-order organization, which itself has signifying potential, are part of their operational semantics; but so also are any disambiguating "incidentals". At all layers, operational semantics may be conditioned or constrained by their context of operation: the "meaning" of a sign is not built into the sign, but is established *ad hoc*, by means of those distinctions between it and possible alternatives that result in particular outcomes in application.

Interestingly, careful consideration suggests that while we often consider that the representational aspect of a sign is fundamental to its operation and provides part of its operational semantics, in general the opposite is the case: operational semantics are primary, and representational semantics can only be established once operational semantics are set and perhaps codified. Once a sign's operation has been established repeatedly it can start to "carry its context with it" — that is, its context can begin to include prior contexts, implicitly recognized — and this is the beginning of representation. (A compelling description of this process in human language development as well as in the more rudimentary linguistic capabilities of chimpanzees, bonobos and gorillas, may be read in [Greenspan and Shankar 2004].) Thus, what any given sign represents must be inferred from its context of operation — which may include the traces of other occasions of its use — and much energy is given, in the application of sign systems to doing useful work in the world, to negotiating these inferences.

The construction of layered sign systems, in which complex representations may be reliably constructed by a rules-based assembly from simpler components, has the precise advantage of managing and reducing this expenditure of energy. Metalanguages (such as an orthography, formal logic, or the grammars of natural or artificial languages), which describe sign systems themselves and stipulate rules for their operation, are more than sterile intellectual exercises. By abstracting and codifying the application of rules to signs, they reduce the need to rely on sheer brute force methods (memorization or negotiation) to support the assembly of signs into sign systems, thus reducing overall complexity and enabling more sophisticated operations at higher levels.

Accordingly, media can be built out of building blocks constituted of other media. Within a (relatively highly evolved) language, utterances take the form of sentences or statements, composed out of words; words are composed of phonemes (or of letters, in the case of written words). When sentences or statements or propositions are combined, they constitute yet another medium at a higher level, which might be identified with argument or narrative. Thus, media are built in layers, each layer a hierarchy of subsystems on a lower layer. This is characteristic of complex systems generally in both the natural and artificial worlds (see [Simon 1996]).

In general, this layering is characterized by two related phenomena:

- The distinctions between the layers becomes clearer over time: media formats evolve and become more systematic and coherent in composition. More and more complex and comprehensive structures become progressively easier to realize, at the cost of a certain kind of expressive power characteristic of early or individual experiments. The higher layers, as they solidify and consolidate, induce a process of simplification and rationalization at lower layers.
- Likewise, as you go up the stack, the distinction between layers becomes less clear. The difference between a letter and a word is almost always clear, but the distinction between a statement and an argument is less so.

As media ascend in complexity, each layer is capable of more "meaningfulness" than the layer it is built on; or at least it has greater representational power, due to the consolidation of operational and (on that basis) even of representational semantics at lower layers. In turn, the context provided by the combination of any set of tokens at any layer provides information relevant to the construing of the local meaning of each of the constituent tokens. This interpenetrating influence between layers, familiar to all students of language and literature, is an important feature of the entire system.

The most complex and "highest" of these layers currently might go by the name of *information architectures*. Until recently we have not really needed to have a name for this layer or members of it (signs that appear in it), as they have for the most part been identified with print media: we have spoken of indexes, cross-references, tables of contexts, abstracts, summaries and bibliographies without being aware that these things might not have to be printed on paper. This is because traditionally, in order to stabilize complex bodies of information, particular features of print media have been indispensible — temporal persistence and asynchrony; "random" or holistic access; conventions of citation; graphic and typographic queues representing organization and providing for navigability; economies of scale afforded by mass production; and so on. Now these features are available in the formats provided by digital media — along with elaborations of them such as the hyperlink, dynamic display and device independence — it becomes useful to consider this higher layer without identifying it with print technology.

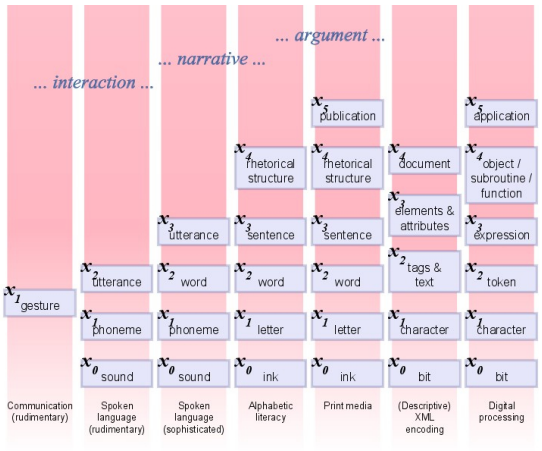


Figure 2: Various different forms of communications media are formed by layering.

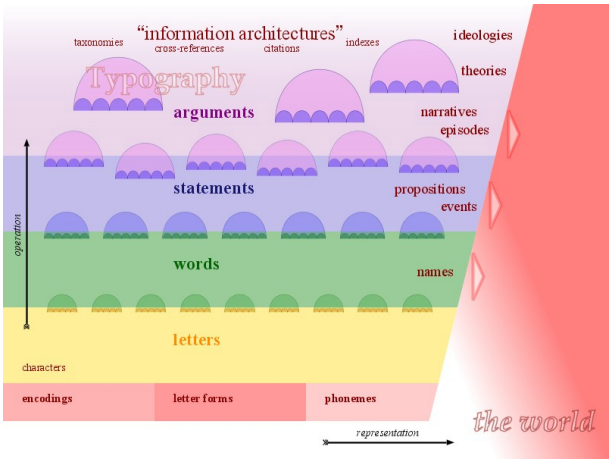


Figure 3: At lower layers, alignments between constituent parts are generally possible, leading to representational correspondences not only between media and the world, but across media themselves. At higher layers, media find their own particular strengths, constituting 'worlds of their own' whose representational capabilities are both presumably greater, and more problematic.

The Emergence of the Digital

Like all media (cf. Marshall McLuhan), digital formats begin as representations (operational refigurations) of prior media. But just as alphabetic literacy, following the logic of its own peculiar operational semantics, quickly assumes forms distinct from the oral forms it begins by mediating, digital media soon become something other than the print counterparts that were to be "transmitted" by telecommunications technologies. Again, it achieves these forms by layering.

Despite the highly developed form of the processing stack when it comes to the computer's own processing (its operations), however, when it comes to digital media we are still at a point

when these layers are inchoate — although the standardization efforts of the last decade (especially as regards HTML, CSS, XML and XSLT) are providing a level of metalinguistic control conducive to their development and maturation. Yet for the most part, applications of digital media are still either derivative of other forms (in the sense that a page on the web may be almost entirely analogous to the same document in print) or directly in service to them; digital media have not fully come into their own.

Nonetheless, the usefulness and power of layering in this context too has long been recognized: we only need to recall the familiar dogma (and the discussion that has surrounded it) of the "separation of format from content" in the design and deployment of markup-based publication systems (see [Sperberg- McQueen and Burnard, 1994], [Durand et al. 1996], [Piez 2001], [Sperberg-McQueen et al. 2002], [Piez 2002]). This tradition recognizes that so-called "descriptive encoding" works by anticipating and expressing at a lower layer (in what we call *source code*), the rationale for structures to be expressed at higher layers through site organization, page layout, typography, screen widgetry, linking and all the apparatus of a full-blown architecture. In this respect, the tagging of an XML document whose encoding is descriptive and data-oriented rather than prescriptive and application-oriented proves to be a connotative system, as the signifiers (the element names "title" and "p") that describe the data (this chunk of text is a nominal paragraph; that one a title) are themselves signs, to be transformed by a heuristic and rules-based process into renditions that will themselves signify to readers that they are paragraphs or titles. This anticipation or prolepsis by markup of further signification elucidates the confusion as to whether we consider descriptive encoding to be at a "higher" or "lower" layer, as indeed it is paradoxically both. Within a classical three-tiered architecture, the XML encoding is "below" its HTML (or print, or SVG, or ODF) rendition; yet we also describe the conversion from descriptive XML into a presentational format such as HTML as a "down-conversion" (since it "loses information"), by rendering only indirectly in presentational features, if at all, what is explicit in the source. The reason we can, in effect, go down to go up, is that here the lower layer achieves its aim of scalability and reusability by working to describe a higher layer that it is not yet practical (if it ever will be) for the computer to infer on its own: it provides, in representational form by a kind of "sleight of hand" (the operational semantics being invisible to the machine and left to the stylesheet designer), information that would ordinarily be available only by a processing context not yet available — the act of reading itself. In fact, the transformation from descriptive XML to HTML is not actually taking it "up" the ladder towards richer information design; it is merely transposing the data into another stack altogether, where, since the operational semantics of HTML are more tightly bound to standard processing contexts (the browser), its implicit functionalities can be elaborated, while the representational aspect can be deferred to where it makes more sense — where, because the tagging now signifies "large and bold", the reader can be trusted to infer "title".

Yet the same discussion has also masked deeper problems and issues (see [Buzzetti 2002], [Caton 2005]) stemming from the limitations in current markup systems, which can gracefully handle only a single organizational hierarchy at a time, and thus lack the representational and expressive power necessary to take full advantage of the computer's capabilities for useful automated processing of complex textual artifacts.

Nevertheless, understanding digital text encoding technologies as complex sign systems elucidates how and why they function without resorting to the metaphysical appeal that "good" encoding should be designed to describe the "thing itself". Especially given the limitations inherent in XML's design, such a position proves soon to be untenable; yet XML systems succeed in doing useful work notwithstanding, and provide the foundations for sustainable, scalable and navigable information resources, whose presentational features (interfaces) can be improved over time. This in itself represents a major advance over what was ever possible in the past.

More generally, a consideration of digital text encoding as a distinctive semiotic system, with its own metalanguages and its own relation to media artifacts, suggests why the humanistic study of digital media remains so foreign to traditional disciplines in the humanities. Likewise, it points the way to the future, as it becomes clear what, and how much, still remains to be done.

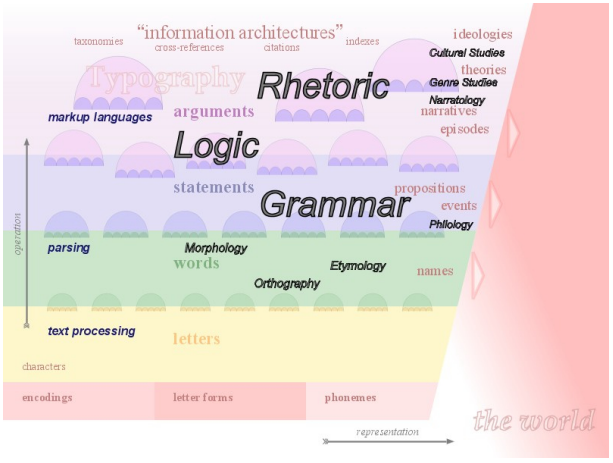


Figure 4: While metalanguages can be expensive to maintain, they provide means not only to describe but also to maintain and control the communications media they describe.

Bibliography

Barthes, Roland. *Elements of Semiology*. Trans. Annette Lavers and Colin Smith. New York: Wang and Hill, 1967.

Buzzetti, Dino. "Digital Representation and the Text Model." *New Literary History* 33.1 (2002): 61-88..

Caton, Paul. "Markup's Current Imbalance." *Markup Languages: Theory and Practice* 3.1 (2001).

Caton, Paul. "LMNL Matters?" *Proceedings of Extreme Markup Languages 2005, Montréal, Québec, August 2005.* 2005. <http://www.idealliance.org/papers/extreme/proceedings/author-pkg/2005/Caton01/EML2005Caton01.zip>

Durand, David, Steven J. DeRose, and Elli Mylonas. "What Should Markup Really Be? Applying Theories of Text to the Design of Markup Systems." *ACH/ALLC 1996.* Available from <http://cs-people.bu.edu/dgd/ach96_talk/Redefining_long.html>

Greenspan, Stanley I., and Stuart G. Shanker. *The First Idea: How Symbols, Language, and Intelligence Evolved from our Primate Ancestors to Modern Humans.* Da Capo Press, 2004.

Maturana, Humberto R., and Francisco Varela. *The Tree of Knowledge: The Biological Roots of Human Understanding.* 1987. Shambhala, 1992.

Piez, Wendell. "Beyond the 'Descriptive vs. Procedural' Distinction." *Proceedings of Extreme Markup Languages 2001, Montréal, Québec, August 2001.* . <http://www.idealliance.org/papers/extreme/proceedings/html/2001/Piez01/EML2001Piez01.html>

Piez, Wendell. "Human and Machine Sign Systems." *Proceedings of Extreme Markup Languages 2002, Montréal, Québec, August 2002.* Ed. B. T. Usdin and S. R. Newcomb. 2002. <http://www.idealliance.org/papers/extreme/proceedings/html/2002/Piez01/EML2002Piez01.html>

Renear, Allen. "The Descriptive/Procedural Distinction is Flawed." *Extreme Markup Languages 2000, Montréal, Québec, August 2000.* Reprinted in Markup Languages: Theory and Practice. 2000.

Renear, Allen H., David Dubin, C. M.Sperberg-McQueen, and Claus Huitfeldt. "Towards a Semantics for XML Markup." *Proceedings of the 2002 ACM Symposium on Document Engineering, McLean, VA, November 2002.* Ed. Richard Furuta, Jonathan I. Malectic and Ethan V. Munson. New York: Association for Computing Machinery, 2002. 119-126.

Saussure, Ferdinand de. *Course in General Linguistics.* Trans. Wade Baskin. 1916. New York: McGraw-Hill, 1966.

Simon, Herbert. *The Sciences of the Artificial.* Cambridge, MA: MIT Press, 1996.

"A Gentle Introduction to SGML." *Guidelines for Electronic Text Encoding and Interchange.* Ed. C. M. Sperberg-McQueen and Lou Burnard. 1994. Chicago: TEI Consortium, 1997. 13-36. <http://www.isgmlug.org/sgmlhelp/g-index.htm>

Sperberg-McQueen, C. M., David Dubin, Claus Huitfeldt, and Allen H. Renear. "Drawing Inferences on the Basis of Markup." *Proceedings of Extreme Markup Languages 2002, Montréal, Québec, August 2002.* Ed. B. T. Usdin and S. R. Newcomb. 2002. <http://www.idealliance.org/papers/extreme/proceedings/html/2002/CMSMcQ01/EML2002CMSMcQ01.html>

Phonemic Accumulations and the Analysis of Poetry

Marc Plamondon *(mplamond@chass.utoronto.ca)*
Nipissing University

I shall present my latest work on the theory of phonemic accumulations and its application to the study of poetry. Computerized stylistics tends to concentrate on words: word frequencies, word co-occurrences, word collocations, and word distributions. My computer-assisted study of stylistics concentrates on the phonemic content of the texts. Computerized phonemic analysis is beginning to yield interesting insights into written texts, especially those, such as poetry, for which sound patterning constitutes a significant element. In *A Companion to Digital Humanities* (2004), Ian Lancashire mentions sound as a future source of inquiry in textual stylistics: "As the implications of cognitive research become understood, text-analysis systems may change, and with them stylistics. Texts in languages whose spelling departs widely from its sounding will likely be routinely translated into a phonological alphabet, before processing, as researchers recognize the primacy of the auditory in mental language processing" ("Cognitive Stylistics and the Literary Imagination"). I have begun undertaking this work with a computer application that translates poems into their broad phonetic transcriptions. The program can then provide visualizations of the phonemic content of the poems, and it can perform calculations on the content.

The theory of phonemic accumulations is based on the theory of the persistence of vision: the effect of a phoneme when reading is carried through to the following phonemes. At the same time, its effect tapers off. When the same phoneme is encountered while the effect of the first phoneme has not been nullified, a cumulative effect of the two phonemes is produced. This gives rise to important sound effects in poetry and literature in general, the most obvious of which is alliteration. The cumulative effect of the /s/ phonemes in "silver silent sails" far exceeds the effect of the individual /s/ phonemes if the words occurred at a greater distance from each other. My computer program is able to quantify the impression the /s/ phonemes may have on the reader of these words.

While many different possibilities present themselves for analysis, my current work focusses on calculations of the fricative accumulations and the plosive accumulations of a poem. The fricative consonants are often regarded as soft sounds, while the plosive consonants are often regarded as harsh sounds. I have already shown that the percentage of plosives and fricatives in poems is mostly constant across poems and authors. Consequently, the occurrence of these phonemes on their own do not produce a significant effect; it is the groupings or occurrences in proximity with each other that seem to produce effects upon the reader. Graphs of their phonemic accumulations exhibit distinct peaks and troughs, and the differences between the two accumulations reveal interesting insights.

For example, Robert Browning's "Two in the Campagna" has a very intense climax of the difference of the fricative and plosive accumulations in its fifth, sixth, and seventh stanzas. These stanzas are those that are most expressive of true love and passion: while the whole poem is an expression of love, more negative thoughts such as the fleetingness of life and the inability to achieve true love on earth are intermingled with the attempts to seduce. The three stanzas with the fricative accumulation climax, however, are the most rhapsodic, the least tainted with doubt. Similarly, in Browning's "Porphyria's Lover," a poem about the murdering of a woman whom the speaker loves too intensely, the stanza (or five-line set—the poem is not officially divided into stanzas) with the greatest climax of the difference of the fricative and plosive accumulations is the fifth, where the speaker of the poem tells how Porphyria expresses her love for him. The rest of the poem is predominantly heavier on the plosive accumulation side. The pattern that emerges is that when the difference between the fricative and plosive accumulations favours the fricatives, the sentiment of the poem is more positive, loving, and sincere. When the plosives are favoured, the sentiment of the poem is more insincere, withdrawn, and bleak. I shall present my results for these and other poems of a similar structure. These are early results, but the results are very promising.

This work is important for the study of literature. This is the first time (that I am aware of) that the phonemic content of a text, i.e. its sound on the page, is put to such advanced calculations. Literary analyses often have to rely upon impressionistic language when discussing the effect of the sounds of the words upon the reader: phrases such as, "the prevalence of s sounds in the final stanza leaves the reader with a soothing, peaceful feeling, one that has countered the chaos of the opening stanzas." Through the analysis of phonemic content such as I am performing, I can provide critics with quantifiable data upon which to base their claims. Further, the possibility that phonemic accumulations are related to the ideas expressed in the poems suggests strongly that a computer can begin to interpret poetry: that it can distinguish passages with expressions of intense, sincere love from passages with expressions of self-doubt or insincerity. Phonemic analysis may produce significant developments in the field of artificial intelligence.

Examples of Images in Text Editing

Dorothy Carr Porter *(dporter@uky.edu)*
University of Kentucky

In 2002, Computers and the Humanities dedicated an issue to "Image-based Humanities Computing" at a time when "a majority of first generation image-based humanities computing projects have reached at least an initial plateau of completion." (Kirschenbaum, p. 3) Since that time interest in incorporating primary source images into "text" editions has blossomed, as can be attested by recent threads on the TEI listserv and work on the TEI council to develop recommendations for specific methods for integrating image files – and pointers to areas of image files – into traditional text encoding projects. The number of image-based projects has multiplied in that time as well, although it takes some effort to find who is working on such projects. There is not (yet) a central listing of all image-based TEI projects under development.

Practical work has been done on tool development since 2002 as well. As of March 1, 2007 there are no less than six tools of which I am aware that serve to edit or display images within the context of text editing. The majority of these tools are for linking text and image for digital display.[1] Add to these additional tools for the simple annotation of images[2] and tools for tagging multimedia.[3] Undoubtedly, the proliferation of tools focused on image editing and display reflects a growing interest in incorporating images into digital editions.

The number of tools available for working with text plus image in digital editing highlights a simple truth: projects and their sources are different, and technologies that will work for one project might be incompatible with another. On the other hand, technologies applicable in simple circumstances might be expanded and combined with other technologies to suit much more complex situations. In this presentation, I will describe the sources of two digital projects with reference to their requirements for becoming viable digital projects. One is quite simple and the other complex, but the same methods inform both projects.

MS Cambridge, Pembroke College MS 25, the subject of the *Digital Edition of Cambridge, Pembroke College MS 25* , (Pembroke 25 project) directed by Paul Szarmach, Director of the Medieval Academy of America, and Thomas N. Hall at the University of Notre Dame. Pembroke 25 is a collection of Anglo-Latin homilies, copied at the scriptorium at Bury St. Edmunds in the late eleventh or early twelfth century by a scribe – or perhaps two scribes – who used the round English Caroline minuscule common there rather than the more pointed Norman Caroline minuscule that came to prominence in England in the period immediately following the Norman Conquest. Following the disillusionment of Bury St. Edmunds in 1538, Pembroke 25 disappeared for a time, but it was given to Pembroke College, Cambridge, at the end of the sixteenth century, and it still lives in that library today. It has been well maintained, it is not damaged, and the script is clear and easy to decipher.

For an edition of this sort, a single text from a single manuscript, the encoding requirements are relatively simple. This manuscript is purely textual, not illustrated or illuminated in any way, but we are noting all abbreviations and distinctive paleographical aspects in the manuscript (including scribal emendations), as well as marginalia. The TEI Header contains some descriptive information, notably a descriptive list of all abbreviation types that are linked to the individual abbreviations throughout the project. We are using the EPPT for the text-image linking in this project, and I will give a brief demonstration of the project as it stands at the time of the conference.

The *Electronic Aelfric* , directed by Aaron Kleist at Biola University and developed by a large group of collaborators, seeks to edit eight Old English homilies by Ælfric of Eynsham, who was arguably the most educated and prolific writer of tenth century England. These homilies cover the period from Easter to Pentecost, and trace their development through six phases of authorial revision and then through nearly 200 years of transmission following Ælfric's death: twenty-four sets of readings or strands of textual tradition found in twenty-eight manuscripts produced in at least five scriptoria between 990 and 1200.

The contrast between the *Electronic Aelfric* and Pembroke 25 is great: while Pembroke 25 is one manuscript, the *Electronic Aelfric* draws from twenty-eight manuscripts. Although no single homily out of the eight occurs in more than ten of these manuscripts, it is still a great number of textual variants to deal with. In addition to the text, the project also needs to address the individual manuscripts – six of which are from the infamous Cotton Collection (now housed in the British Library), damaged by fire in 1731. Those manuscripts that are not damaged still have singularities, such as marginalia, that we also wish to encode and link to image. We are using the EPPT for this project, partnered with the TEI Apparatus tags, to bring together the text and images of several different manuscripts. I will show examples of corresponding manuscript pages, as well as sample code illustrating multiple variants partnered with image-text linking.

1. Edition Production and Presentation Technology (EPPT), developed by Kevin Kiernan under the aegis of the *Electronic Boethius* project, <http://www.eppt.org/eppt-trial/EPPT-TrialProjects.htm>
   Image processing services, developed by Neel Smith and Christopher Blackwell through Harvard's Center for Hellenic Studies, at *Digital incunabula: a CHS site devoted to the cultivation of digital arts and letters*, <http://chs75.harvard.edu/projects/diginc/techpub/images>
   Juxta, developed through the NINES project (networked infrastructure for nineteenth-century electronic scholarship), <http://www.nines.org/tools/juxta.html>
   Florian Thienel, “Konzept für einen editionsphilologischen EDV-Arbeitsplatz auf der Basis von XML und verwandten Standards” Diplomarbeit im Fach Informatik, Universität Würzburg
2. UVic Image Markup Tool (1.3.0.3), <http://www.tapor.uvic.ca/~mholmes/image_markup/>
3. Doug Reside at the Maryland Institute of Technology in the Humanities (MITH) at the University of Maryland is developing still-unnamed tool to tag not only images, but video and audio files as well

Bibliography

Carlquist, Jonas. ""Medieval Manuscripts, Hypertext and Reading. Visions of Digital Editions." *Literary & Linguistic Computing* 19.1 (2004): 105-118.

Cross, James E. *Cambridge Pembroke College MS. 25: A Carolingian Sermonary Used by Anglo-Saxon Preachers.* King’s College London Medieval Studies, 1. London: King's College, 1987.

Dué, Casey, and Mary Ebbott . "As Many Homers As You Please: an On-line Multitext of Homer." *Classics @* (2004). <http://classics.furman.edu/classicsAt2/due-ebbott_2004_all.html>

Godden, Malcolm. *Aelfric’s Catholic Homilies: Introduction, Commentary and Glossary*. New York: Published for the Early English Text Society by the Oxford University Press, 2000.

Kiernan, Kevin. "Digital Facsimiles in Editing: Some Guidelines for Editors of Image-based Scholarly Editions." *Electronic Textual Editing.* Ed. Lou Burnard, , Katherine O’Brien O’Keeffe and John Unsworth. New York: Modern Language Association, 2005. preprint at <http://www.tei-c.org/Activities/ETE/Preview/kiernan.xml>

Kiernan, Kevin, Jerzy W. Jaromczyk, Alex Dekhtyar, and Dorothy Carr Porter. "The ARCHway Project: Architecture for Research in Computing for Humanities through Research, Teaching, and Learning." *Literary & Linguistic Computing* 20 (Suppl 1) (2005): 69-88. doi:10.1093/llc/fqi018

Kiernan, Kevin, W. Brent Seales, and James Griffioen. "The Reappearances of St. Basil the Great in British Library MS Cotton Otho B. x." *Computers and the Humanities* 36.1 (2002): 7-26.

Kirschenbaum, Matthew G. "Editor's Introduction: Image-based Humanities Computing." *Computers and the Humanities* 36.1 (2002): 3-6.

Lecolinet, Eric, Laurent Robert, and Francois Role. "Text-image Coupling for Editing Literary Sources." *Computers and the Humanities* 36.1 (2002): . 49-73.

ACH Employment Committee

Stephen Ramsay *(sramsay@unlserve.unl.edu)*
University of Nebraska

Members of the ACH Employment Committee will be available to answer questions about career planning, employment opportunities, and the state of the job market in digital humanities.

(Offered in conjunction with the ACH society panel, "Employment - Pedagogy - Professionalisation.")

Digital Text Resources for the Humanities – Legal Issues

Georg Rehm *(georg.rehm@uni-tuebingen.de)*
Tübingen University
Andreas Witt *(andreas.witt@uni-tuebingen.de)*
Tübingen University
Erhard Hinrichs *(eh@sfs.uni-tuebingen.de)*
Tübingen University
Timm Lehmberg
(timm.lehmberg@uni-hamburg.de)
Hamburg University
Christian Chiarcos
(chiarcos@ling.uni-potsdam.de)
Potsdam University
Felix Zimmermann *(mail@felix-zimmermann.eu)*
Institute for Legal Informatics
Hannover University
Heike Zinsmeister
(heike.zinsmeister@uni-tuebingen.de)
Tübingen University
Johannes Dellert *(jdellert@sfs.uni-tuebingen.de)*
Tübingen University

The session "Digital Text Resources for the Humanities – Legal Issues" consists of three papers that address the legal aspects connected to several crucial phases of handling text resources: collecting, compiling, curating, analysing, distributing, and archiving text resources such as corpora, are tasks carried out on a day-to-day basis by people involved in fields such as, for example, humanities computing, computational and corpus linguistics, information retrieval and text mining. Despite the ubiquity of document collections, the legal issues that are intrinsically tied to virtually all texts created and published by third parties (most importantly, their copyright, as well as privacy issues), do not typically attract a lot of interest. Though these issues are acknowledged, they are often regarded as rather insignificant for the research question at hand, or a project does not have any jurisprudential expertise to deal with legal issues in an adequate way. As a consequence, distributing a corpus (for example, to other interested

researchers) whose provenance is unknown or questionable, or publishing excerpts from a document collection on a website, may become next to impossible from a legal point of view. This is why scholars often decide not to publish their collections (or parts thereof) online at all, in order to avoid any potential legal problems. The session aims to provide an overview of the following legal aspects:

- The first contribution, "Language Corpora – Copyright – Data Protection: The Legal Point of View" (Timm Lehmberg, and Felix Zimmermann), highlights the legal requirements that hold with regard to the construction of digital text resources, special emphasis is given to the aspect of copyright and data protection (for example, potential reasons for the need to anonymise text corpora).
- The second presentation, "Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System" (Timm Lehmberg, Christian Chiarcos, Erhard Hinrichs, Georg Rehm, and Andreas Witt), consists of two parts: first, a web-based questionnaire is introduced that was developed to capture the requirements research projects have with regard to the archiving and distribution of their corpora; second, initial results from a study that spans three large research centres and more than 60 individual research projects are reported.
- The final paper, "Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections" (Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert), introduces the idea of masking an annotated text corpus whose original source text collection is copyright-protected, so that the masked version can be distributed without any restrictions; furthermore, a fully working tool for masking an XML-annotated corpus is presented.

The authors of the three papers are associated with a joint project situated in three Collaborative Research Centres (SFB, Sonderforschungsbereich) that are sponsored by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft): SFB 441 (*Linguistic Data Structures*, Tübingen University), SFB 538 (*Multilingualism*, Hamburg University), and SFB 632 (*Information Structure*, Potsdam University). Each of these three research centres consists of about 15 to 20 research projects. Most projects work with digital text collections, in practically all cases these collections and corpora are constructed by the respective researchers themselves. A problem people involved in the fields of digital humanities or computational linguistics are often confronted with concerns the fact that the sustainability and reusability of corpora is not given too much attention – or that these aspects, in a worst case scenario, are completely ignored. Corpora are often created for an application or for a project that has a very specific research question, but when the project is finished it becomes next to impossible (especially for third parties) to gain access to the resource that took several months or maybe even years to create. The joint project *Sustainability of Linguistic Data* was therefore established to provide the conceptual, technical and infrastructural basis for a solution to the problem of sustainably archiving these digital text collections, addressing issues as diverse as, for example, annotation and metadata frameworks, best practice guidelines, legal issues of distributing text collections, and unifying diverse tag sets by means of an ontology.

Session Chairs: Georg Rehm and Andreas Witt

Language Corpora – Copyright – Data Protection: The Legal Point of View

Felix Zimmermann and Timm Lehmberg

1 Introduction

Creating comprehensive and sustainable archives of linguistic data and making them (or parts of them) accessible to the research community leads to a number of essential legal questions being raised by different aspects of law. Like any discipline handling large amounts of data, the digital humanities are confronted with a complex system of authorities and restrictions. From acquisition, through storing and processing to the annotation and finally publication of the data, there are a number of rights as well as duties each participant in this process has to consider. Additionally, some legal systems provide special rules for the use of data for scientific purposes. On the one hand the opacity of the legal position leads to the assumption that, in many cases, linguistic data are used and transferred in a way that does not comply with legal requirements. On the other hand there is a noticeable tendency not to transfer linguistic data for fear of breaking the law (see Jüttner 2000, and Patzelt 2003).

2 Relevant Areas of Law

Two different areas of law play an important role in the use of linguistic data for research purposes:

- "*Intellectual Property Rights*" provide legal protection of non-material goods which are any kind of intellectual property of a third party. This includes, amongst others, literary works as well as *databases*, software and utility patents. In terms of law language corpora are defined as databases.
- "*Privacy and Personal Data Protection Law*" imposes restrictions for the processing of any personal data, i.e., any data that can be linked to an individual. In the face of linguistic data processing any audio and video recordings

(and their transcriptions) as well as metadata that contain personal information on speakers are covered by this law.

Both areas are relevant to the complete process of data processing and have to be considered from the initial step of the data based work (normally the acquisition of the data) to the time of publication.

3 Aspects of National and International Law

In everyday legal practice a particularly relevant role is played by those legislative rulesets that are based on constitutional norms. Within these, interests and entitlements of other involved individuals and institutions, which are worthy of protection, are often outlined in minute detail in relation to the procurement, processing, and transfer of linguistic primary data.

Federal states, which contain individual member states with their own legislative authority (such as the US, Germany, Switzerland, Austria, Spain) may have enacted specific member state rules. This leads to the possibility that there may be complex and potentially internally conflicting legislation within a state in a federation.

It is not just, however, the original national legal situation which regulates the use of linguistic data. International obligations may, through direct or indirect applicability, have considerable impact. In 2007, 27 member states of the European Union adhere to European legal instruments (such as directives and regulations) in relation to the national and international use of data. Pursuant to the doctrine of direct applicability enshrined in Article 10 of the Treaty establishing the European Communities, these norms have priority in relation to potentially conflicting national norms. What needs to be borne in mind is that the individual member states have some leeway in the implementation of the instruments, which may lead to minute differences in the level of protection.

Finally, public international treaties which oblige their signatories to adhere to certain minimal standards need to be taken into consideration. In relation to linguistic data and the problem of copyright, the *Copyright Treaty of the World Intellectual Property Organisation* (*WIPO*, 1996) is to be considered as particularly relevant. The question of personality rights with a view to individuals whose data are processed is addressed in the *Convention on Human Rights and Fundamental Freedoms* (1950). Additionally, the *Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data* (1981) provides further normative guidance for the member states of the European Union.

4 The Legal Impact of Intellectual Property

Copyright protection of language corpora is provided by different aspects of applicable law. In order to simplify the presentation, there will be a focus on the law of harmonised rules by the European Communities that are placed within the framework of the World Intellectual Property Organisation (WIPO).

4.1 Directive 91/250/EC on the Legal Protection of Computer Programs

The different tasks of linguistic data processing (transcription as well as annotation etc.) require a considerable number of software tools. For this purpose, apart from commercial development, software is written by the research establishment's employees. The participants in this process rarely bother with legal protection of their work. By implementing the Directive 91/250/EC, computer programs in all Member States of the European Community are protected by copyright law. In accordance with Article 1.3 of the Directive 91/250/EC, a computer program is protected, if it is original in the sense that it is the author's own intellectual creation. Ideas and principles of a computer program are not protected by this Directive. The term of protection is the author's lifetime plus a period of 50 years. The author owns the exclusive rights to reproduce, translate, adapt and publicly distribute his computer program.

If a computer program has been created by an employee, in accordance with Article 2.3 of the Directive 91/250/EC, the employer is, unless otherwise provided by contract, the copyright holder of the resource. In the case of software being developed within a research project, from this point of view the copyright is held by the respective research establishment (University etc.).

4.2 Directive 96/9/EC on the Legal Protection of Databases

In accordance with Article 1.2 of the Directive 96/9/EC, a database is defined as a "collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means". Without exception, linguistic corpus data come under this protection. This Directive makes two significant stipulations. First, it offers protection by copyright to databases which, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation. Thereby the author owns the exclusive right to carry out or authorise the reproduction, alteration and distribution. Secondly the Directive creates an exclusive right protection *sui generis* for makers of databases, independent of the degree of innovation. This protection of any investment allows the makers of databases to prevent unauthorised extraction and/or re-utilisation.

4.3 Copyright Directive 2001/29/EC

The Copyright Directive 2001/29/EC adapts legislation on copyright and related rights to reflect technological developments into European Community law. In this process, it discusses and harmonises the property of reproduction, communication and distribution rights. Concerning linguistic research data, attention should be paid to Article 5.3(a) of the

Copyright Directive. It gives freedom to Member States in supporting non-commercial science by making copyright less restrictive for academic use of copyrighted work.

5 The Legal Impact of Data Protection

Directive 95/46/EC on the protection of individuals with regard to the processing of personal data imposes strict restrictions for the elevation and utilisation of personal data. Personal data are pieces of information which can be linked to a specific person. The processing of personal data only is permitted by law, if there is a clear and lawful purpose at the time of data procurement, and if the respective person has expressed his/her consent. Further restrictions are imposed, if the racial, national or ethnical origin, political opinion, religious or philosophical beliefs are apparent. The same applies to the disclosure of health conditions or sexual life. If personal data are transferred to countries outside of the European Union (Transborder Dataflow to third countries), a level of protection has to be guaranteed that is equivalent to the European level, for example by means of the *Safe-Harbour-Principles*. The respective person may enforce his/her rights by means such as disclosure and deletion of the data. Article 6.2, Article 11.2 and Article 13.2 of the Data Protection Directive contain privileges for academic research. An escape strategy in respect of data protection law problems is complete anonymisation (disguising by removing personal information by abbreviating names, locations etc.) or pseudonymisation (disguising by aliasing individuals, locations, etc.) of the personal data. However, it remains unsolved which level of abstraction constitutes sufficient anonymisation, particularly if it is possible to draw conclusions by joining the data with other resources.

Figure 1 gives an overview about the different types of right holders to a database.

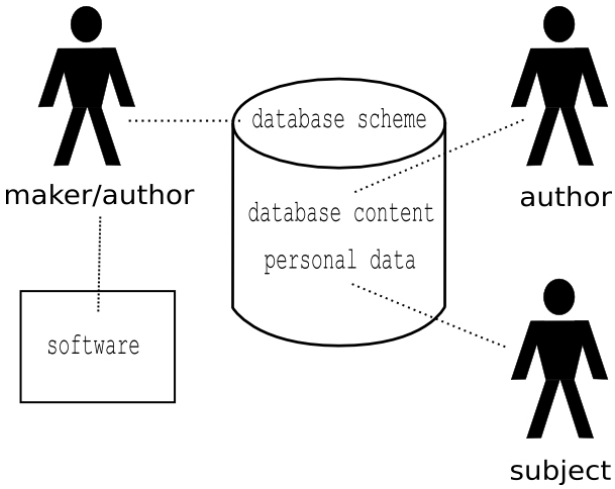


Figure 1: The different types of right holders

Legal Competence by Trusted Third Parties

An additional option is given by the use of a trusted third party hosting the information that has been disguised by anonymisation or pseudonymisation. It may act as a trustee, passing the aliased or anonymised data from its origin to a requesting research institution. The trusted party is not required by law, but it has the ability to provide a high level of data security, integrity and protection during the whole data transaction process (Kilian et al 1995, p. 63). Additionally a trusted party can provide specialist advice in technical and copyright matters. Further, we suggest proceedings to increase legal certainty in case of creating and using linguistic databases.

Bibliography

Jüttner, Irmtraud. "Mannheimer Korpus und Urheberrecht. Die Einbeziehung zeitgenössischer digitalisierter Texte in die computergespeicherten Korpora des IDS und ihre juristischen Grundlagen." *Sprachreport* 3 (2000).

Kilian, Wolfgang. "Daten für die Forschung im Gesundheitswesen." *Gutachten II.* Toeche-Mittler Verlag, 1995. 57-76.

Patzelt, Johannes. "Unter juristischem Blickwinkel: Textkorpora und Urheberrecht." *Korpuslinguistik deutsch: synchron – diachron – kontrastiv: Würzburger Kolloqium 2003,*. Ed. Werner Wegstein and Johannes Schwitalla. Würzburg, 2003.

Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System

Timm Lehmberg, Christian Chiarcos, Erhard Hinrichs, Georg Rehm, and Andreas Witt

1 Introduction and Overall Concept

Most metadata standards used for corpus linguistic purposes (TEI, OLAC, IMDI etc., for a complete overview see Lehmberg and Wörner 2007) require elements that contain legal information about the rights holder to the particular resource and/or its accessibility. Normally these metadata elements are kept very abstract and do neither distinguish between the different types of personal rights nor do they consider the option of multiple holders of copyright.

The legal situation upon which the evaluation of linguistic data to be used for scientific purposes is based is clearly defined, but too complex to be understood completely by non-experts. Furthermore, it varies from one country to the other and is in a constant state of flux.

In the framework of our joint sustainability initiative (see the introduction to this session), a large number of heterogeneous corpora have been acquired from multiple sources and multiple projects, and processed with regard to different individual requirements (Schmidt et al. 2006). This heterogeneity is responsible for the problem that the legal metadata that need to be collected strongly vary with regard to the respective corpus and data situation. Only for a small number of projects associated with our sustainability initiative are detailed sets of legal metadata that inform a potential user of the corpus about, for example, stipulations or copyright holders, readily available. For the majority of projects and corpora, this task has to be performed retroactively.

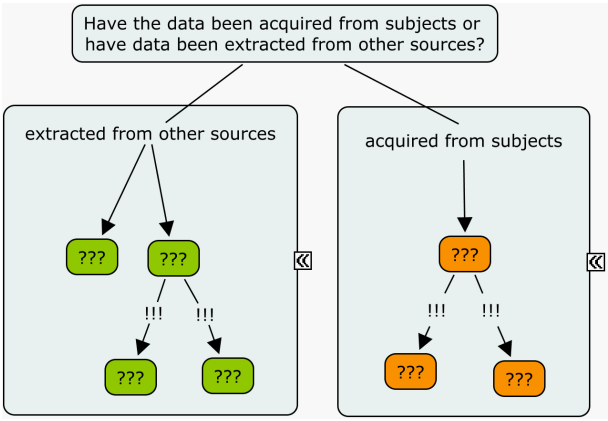


Figure 1: A concept map visualising the query structure

Facing the complexity of the legal context (see Zimmermann and Lehmberg, in this session), it is almost impossible for non-experts to evaluate the situation of their language data and to extract the relevant metadata without professional advice.

To reduce the complexity of this task, concept maps were created with the goal of making the legal situation as well as the legal terminology transparent and understandable to non-professionals. Unlike mindmaps that are primarily used for the (often spontaneous and intuitive) mapping of ideas and processes, the technique of concept mapping is intended more for knowledge modelling: concepts are represented by nodes, links represent the relations between them.

As a utility to create the concept maps modelling the legal situation within our joint sustainability initiative we used *CmapTools*, a program distributed by the Institute for Human and Machine Cognition (IHMC). IHMC CmapTools provide a client/server architecture that allows users at different locations to work collectively on Concept Maps and to discuss their structure and content online.

Based on these schemata and following the principles of decision-trees, we built an additional concept map representing the query structure of a questionnaire. Digressing from the original principles of concept mapping mentioned above, in this map queries are represented as nodes whereas responses are represented as links between them. The primary query given in the centre node (see figure 1) corresponds to two central aspects of law (data protection and copyright, see Zimmermann and Lehmberg, in this session). Each response leads to a large number of additional queries that again, depending on the users' response, have subordinated queries. Further sections of the concept map deal with the accessibility of the data as well as their respective principles and standards of data processing.

In same manner we modelled the query structure that surveys the meta information that ideally has been collected in connection with the compilation process of corpora. Therefore it contains queries asking for established metadata standards (TEI, DC, OLAC, IMDI etc.) that may have been used, and if necessary asks for additional information.

Due to the fact that the IHMC CmapTools provide an export of concept maps into an XML-based format, the content and structure of the concept map can be processed automatically to create the web based questionnaire that is described in the following section.

The complete concept map structures will be demonstrated in conjunction with example scenarios in our presentation.

2 Implementation

As the questionnaire has to be accessible from different research project locations, it has been implemented using a XAMP (any operating system, Apache, MySQL and PHP) architecture to create a user-friendly, web-based interface. The conceptual structure represented by the concept map is transformed into a relational database model. Accordingly, it is possible both to model the tree structure of the queries (Celko, 2004) and to save responses to these questions within the database. Additionally, the database includes user data (as well as user access control data) and links them to the metadata sets of the resource being acquired by the questionnaire.

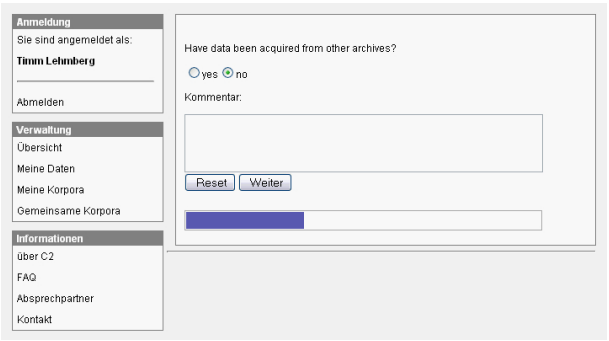


Figure 2: A web-based wizard guides the user through the questionnaire

The user interface is generated by a script that parses the database and guides the user though the questionnaire tree with

the help of a web-based wizard (see figure 2). This architecture has several advantages:

Figure 3: An overview page gives information about the data collection progress

- Subordinate queries that refer to specific details of some legal aspects automatically can be skipped if they become superfluous. For instance, there is no need to query contractual agreements with subjects if there is no personal data contained in the corpus.
- The data model provides users with the option of registering multiple corpora and running the questionnaire wizard individually. Furthermore, users can share the data they entered into the system with other registered users so that it is possible to edit the data across project locations (for example, queries can be skipped, answered later, or left to other users).
- Should the structure or content of the questionnaire tree be changed, the database will be modified accordingly. If the change leads to unanswered queries, this will be indicated to the user in a status page. For this reason, every user account has an overview page that gives information about the state of progress of every registered resource (see figure 3).
- The questionnaire includes queries about metadata content and standards that already have been applied to the registered corpora, so that users do not have to insert redundant information already contained in existing metadata sets.
- Administrator users have unlimited access to all data in the database, so that users can be provided with support, if needed.

We are currently in the process of collecting legally relevant metadata from about 60 different research projects with the aid of the questionnaire system described in this paper. Content and structure of the concept maps is available on our project homepage at `<http://www.sfb441.uni-tuebingen.de/c2/>`.

Bibliography

Celko, Joe. *Trees and Hierarchies in SQL for Smarties.* San Mateo: Morgan Kaufmann, 2004.

Lehmberg, Timm, and Kai Wörner. "Annotation Standards." *Corpus Linguistics, Handbücher zur Sprach- und Kommunikationswissenschaft (HSK).* Ed. Anke Lüdeling and Merja Kytö. Berlin: de Gruyter, In press.

Schmidt, Thomas, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. "Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources." *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art, East Lansing, Michigan, June 2006.* 2006.

Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections

Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert

1 Introduction

Though XML-annotated text collections are commonplace in humanities computing, the value of the annotation is often underestimated, as interesting applications can be realised by ignoring the content and considering the annotation exclusively. At the same time, the distribution of text collections (e. g., linguistic resources) is often restricted by rigid licence agreements. Usually, a corpus consists of a source text collection (STC) acquired from third parties such as web sites or publishers, and annotation layers that refer to, for example, structural or linguistic properties. In practically all cases the STC is a copyrighted property, so that it is up to the copyright holder to decide if, and under which conditions, the corpus - a crucial part of which is the STC - can be made available to the public or to the research community.

The example we use in this paper is TüBa-D/Z ("Tübingen Treebank of Written German" (Telljohann et al, 2004 & Telljohann et al, 2006)). This manually annotated treebank is based on a CD ROM that contains an archive of the issues the newspaper *die tageszeitung (taz)* has published since 1986. If a researcher (the licencee) wants to obtain TüBa-D/Z, available for academic purposes free of charge, he or she has to sign a licence agreement with Tübingen University's Linguistics Department (the licencer) which states that the licencer is the copyright holder of the annotation and that the STC, as published on the *taz* CD ROM, is copyrighted by contrapress

media GmbH. The licencee has to certify that he, she or the institution the person works for has a valid licence for this CD ROM.[1]

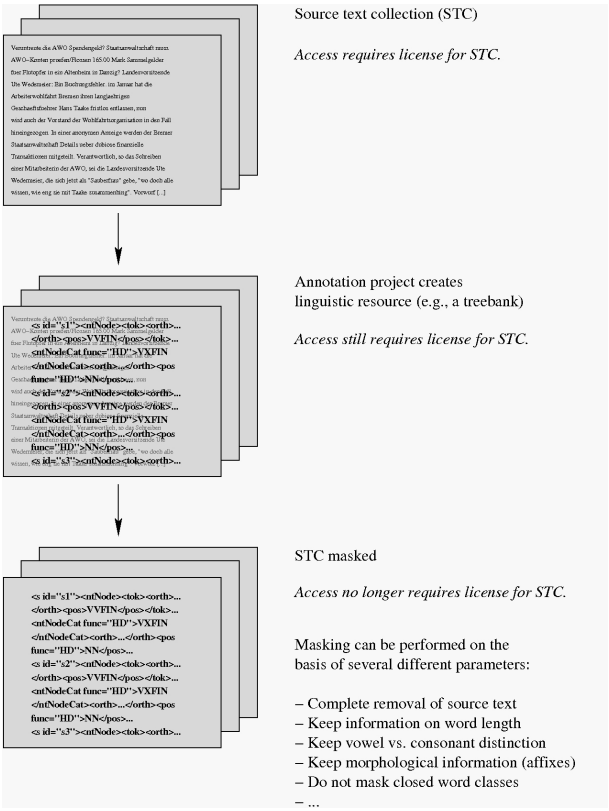


Figure 1: Masking linguistic corpora by example of the TüBa-D/Z treebank

We propose the notion of corpus masking, i. e., obfuscating the STC, but not the annotation layer(s), the STC is "removed", so that the original licensing restrictions no longer hold for the "new" resource. The advantage is that the valuable annotation information can be made available for free (see figure 1).[2]

2 Corpora – Licence Restrictions – Sustainability

When linguists have created a corpus it can become quite difficult to gain access to the corpus once the project is finished. In an ideal world, academics can turn to a sustainability initiative in order to archive their datasets and to make them available to other researchers, e. g., by means of a web-based corpus platform (Dipper et al, 2006 & Schmidt et al 2006). Apart from issues such as providing standardised markup languages and metadata sets (Chiarcos et al, 2006 & Wörner et al, 2006), sustainability initiatives have to take the copyright of the original data into account.

We developed a tool that is able to mask corpora on the fly. Should someone who is interested in a corpus that is available under a rigid licence model not have a valid STC licence, he or she can still receive the corpus, albeit in masked form. A corpus potentially can be associated with *several* accessibility regulations: full access to TüBa-D/Z requires a licence for the *taz* CD ROM, whereas masked versions can be placed under, say, the GNU Free Documentation or a Creative Commons Licence. Therefore, a sustainability initiative has to come up with a flexible system of representing the relationships and dependencies between the STC and the different annotation layers and their individual licence restrictions.

3 How to Mask Linguistic Resources

The easiest option to obfuscate an annotated corpus is to remove the text. A less radical solution substitutes every STC character with, for example, "x" and every digit with "0". In addition to preserving word length, this process retains information on upper and lower case by substituting these with "x" and "X" (Toms & Campbell, 1999).

We developed *CorpusMasker*, a Java-based tool for the parameterised masking of linguistic resources represented as XML documents. The XML element(s) or attribute(s) that comprise the actual words or tokens to be masked (in case of TüBa-D/Z, the `<orth>` element) can be specified to handle arbitrary annotation schemes. CorpusMasker features a dictionary approach: after collecting all word forms, every word is mapped onto a randomly generated string and replaced by that string. Word length can be retained, as well as information on the distribution and positioning of vowels and consonants. If a word is usually written with an initial lower case character and that word appears with an initial upper case character, the same randomised word is used (e. g., "dort" -> "kulp", "Dort" -> "Kulp"). CorpusMasker performs an affix analysis that is similar to morphology induction. The algorithm analyses certain words, masks the roots, but retains the affixes, so that the text is masked but valuable linguistic information that in itself is insufficient to reconstruct the source text or even to interpret the masked text, is kept intact for further analysis. Parameterised masking can be performed with several different degrees of retaining linguistic information, from the complete removal of the STC to a rather light but sufficient masking that keeps, e. g., closed word classes unchanged (see table 1; affixes are marked in italics).[3]

Linguistic corpora often contain POS information so that the randomisation process results in a list that could act as a key to unlock the masked corpus, i. e., to reconstruct the STC. As publication of this complete list would contradict the purpose of the tool, we will only provide a reduced version of the file so that the randomly generated words can be mapped onto POS tags.

Part-of-speech (POS):	VVFIN	ART	NN	NN	
Original sentence:	Veruntreute	die	AWO	Spendengeld	?
Characters replaced with [xX9]:	Xxxxxxxxxx	xxx	XXX	Xxxxxxxxxx	?
Random characters:	Sololplaoka	tao	UJA	Wkirdomgirk	?
Random characters, keep affixes, keep closed word classes:	Verildniite	die	AJE	Storparpamb	?

Table 1: Masking examples for "Veruntreute die AWO Spendengeld?"

4 Masked Corpora: What are They Good for?

Our original goal had been to give researchers interested in TüBa-D/Z the option of examining the annotation without ordering the *die tageszeitung* CD ROM first. As our sustainability platform will give access to copyrighted corpora, we will implement the option of masking a corpus archive before every single download to enhance security. Furthermore, a password protected dictionary lookup could be provided that enables researchers to retrieve a small amount of translations from randomised strings back to original words. Following, we sketch some application scenarios for masked corpora.

Unlexicalised parsing A masked corpus can be used for all sorts of unlexicalised training. Charniak (1996) shows that an unlexicalised PCFG trained on treebank annotations is compatible with other unlexicalised parsers. In addition to the masked training data, a minimal amount of testing data was required. In the case of TüBa-D/Z this subcorpus could consist of randomly shuffled example sentences from the treebank with unmasked text and full annotation. Hinrichs et al. (2005) discuss experiments in memory-based learning of anaphora resolution. Their tool is trained on the annotation of TüBa-D/Z and does not take lexical information into account. The features refer to morphological properties, parts-of-speech, syntactic boundaries and grammatical functions, all of which are available in the annotation. In this case even the test data could be generated directly from the masked resource since the annotation includes marking of equivalence classes comprising pronouns and noun phrases. The gold standard for testing consists of these equivalence classes only in which the words are represented by positional indices. The evaluation would then test whether the relevant indices are grouped together correctly. A comparable tool trained on masked corpus data could as well be applied to `real' German texts.

Qualitative and quantitative analyses TüBa-D/Z's annotation can be used for qualitative and quantitative analyses, it includes both syntactic categories as well as grammatical functions. A linguist can, for example, examine which categories occur as predicatives (element PRED). In addition to this qualitative investigation, the corpus also allows a quantitative analysis: what percentage of predicatives is realised by a noun phrase, what percentage is realised by an adjectival phrase or by a prepositional phrase? To give a second example, coordinate structures are marked with the label KONJ; even without knowledge of the word level the treebank annotation gives sufficent information to examine parallelism effects with respect to the structure of the conjuncts: syntactic categories, grammatical functions, modifiers, and length, see, e. g., Levy (2004), and Steiner (2006).

Teaching linguistics and computational linguistics The masked version of TüBa-D/Z contains an unnatural language that acts like German syntaxwise, but the lexicon of this language contains, for the most part, random strings and associated POS tags. This fact makes the masked treebank a valuable resource in the context of teaching computational linguistics. If students have to work with a language that has a known syntax and a rudimentary morphology but lexical entries that bear no meaning whatsoever, they might be able to concentrate better on the tasks of developing grammar rules or improving parsing efficiency (e. g., with regard to unlexicalised parsing). This approach of blanking out semantics is compatible with Chomsky's notion of language as processing a set of symbols.[4]

Evaluating NLP software Another promising application scenario is the evaluation of NLP software. Most tools use n-gram language models, more sophisticated applications can be trained on annotated corpora. With a masked resource it is possible to measure the influence syntactic annotations have concerning precision and recall, as the performance data of an NLP tool with regard to original, as well as slightly and fully masked corpora can be compared. This approach could result in substantial arguments in favour, or against the use of treebanks for training NLP tools.

5 Related Work

Anonymisation methods remove proper nouns and other identity-revealing phrases to protect the privacy of the people mentioned in a text (for example, medical or legal records (Corti et al, 2006, Medlock, 2006, Poesio et al, 2006, & Rock, 2001)). A second application area is concerned with the removal of cues that might reveal the identity of the author of a text. A third area concerns the masking, or obfuscation of texts, as described in the present paper; we are not aware of similar approaches to the masking of linguistic resources.[5]

6 Concluding Remarks

We call our approach parameterised masking because the randomisation process can be influenced with regard to several parameters, so that, for example, certain word classes are not randomised. Typically, when closed word classes such as determiners and prepositions are kept intact, at least part of the original meaning of a sentence can be guessed. This leads us to a crucial question: what happens if we choose to mask only a small number of words (for example, only proper nouns)?

Do we have to mask a certain percentage of words, in order to bypass the STC's licensing restrictions? When does a text that has been masked only minimally become the original text again, so that the licence restrictions *prohibited* the distribution of the pseudo-masked linguistic resource?

Acknowledgements

The authors would like to thank Timm Lehmberg (Hamburg) and Felix Zimmermann (Hannover) for valuable comments with regard to legal aspects of the approach presented in this contribution. Furthermore, we would like to thank our colleague Holger Wunsch for valuable discussions.

Bibliography

Charniak, E. "Tree-bank Grammars." *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96).* MIT Press, 1996. 1031-1036.

Chiarcos, Christian, Timm Lehmberg, Georg Rehm, and Andreas Witt. *Regulating Access to the Sustainability Platform. Technical report.* SFB 441 (Tübingen University), 2006.

Corti, L., A. Day, and G. Backhouse. "Confidentiality and Informed Consent: Issues for Consideration in the Preservation of and Provision of Access to Qualitative Data Archives." *Forum: Qualitative Social Research* 1.3 (2000).

Dipper, S., E. Hinrichs, T. Schmidt, A. Wagner, and A. Witt. "Sustainability of Linguistic Resources." *Proceedings of the LREC 2006 Workshop Merging and Layering Linguistic Information, Genoa, Italy.* Ed. E. Hinrichs, N. Ide, M. Palmer and J. Pustejovsky. 2006. 48-54.

Hinrichs, E., K. Filippova, and H. Wunsch. "What Treebanks Can Do For You: Rule-based and Machine-learning Approaches to Anaphora Resolution in German." *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT 2005), Barcelona, Spain.* Ed. M. Civit, S. Kübler and Ma. Antònia Martí. 2005. 77-88.

Levy, Roger. Presented at the Department of Linguistics, University of Colorado-Boulder, March 11, 2004. 2004.

Medlock, B. "An Introduction to NLP-based Textual Anonymisation." *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.* Ed. N. Calzolari, K. Choukri, A. Gangemi, J. Mariani, B. Maegaard, J. Odjik and D. Tapias. 2006. 1051-1056.

Piez, Wendell. "Way Beyond Powerpoint: XML-driven SVG for Presentations." *Proceedings of XML 2004, Washington, November 2004. IDEA.* Ed. Lauren Wood. 2004.

Poesio, M., M. A. Kabadjov, P. Goux, U. Kruschwitz, E. Bishop, and L. Corti. "An Anaphora Resolution-Based Anonymization Module." *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.* Ed. N. Calzolari, K. Choukri, A. Gangemi, J. Mariani, B. Maegaard, J. Odjik and D. Tapias. 2006. 1191-1193.

Rock, F. "Policy and Practice in the Anonymisation of Linguistic Data." *International Journal of Corpus Linguistics* 6.1 (2001): 1-26.

Schmidt, Thomas, Christian Chiarcos, Timm Lehmberg, Georg Rehm, Andreas Witt, and Erhard Hinrichs. "Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources." *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art, East Lansing, Michigan.* 2006.

Steiner, Ilona. "Coordinate Structures: On the Relationship between Parsing Preferences and Corpus Frequencies." *Pre-Proceedings of the International Conference on Linguistic Evidence 2006, Tübingen, February 2006.* 2006.

Telljohann, Heike, Erhard Hinrichs, Sandra Kübler, and Heike Zinsmeister. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report.* Tübingen: Seminar für Sprachwissenschaft, Universität Tübingen, 2006.

Telljohann, Heinke, Erhard Hinrichs, and Sandra Kübler. " The TüBa-D/Z Treebank – Annotating German with a Context-Free Backbone." *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, 2004.* 2004.

Toms, E. G., and D. G. Campbell. " Genre as Interface Metaphor: Exploiting Form and Function in Digital Environments." *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32).* IEEE Computer Society, 1999.

Varga, D., P. Halácsy, A. Kornai, V. Nagy,, L. Németh, and V. Trón. "Parallel Corpora for Medium Density Languages." *International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria, 2005.* Ed. G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov and N. Nikolov. 2005. 590-596.

Wörner, Kai, Andreas Witt, Georg Rehm, and Stefanie Dipper. "Modelling Linguistic Data Structures." *Proceedings of Extreme Markup Languages 2006, Montréal, Québec, August 2006.* Ed. B. T. Usdin. 2006.

1. The *taz* CD ROM costs about 50 Euros. Licences for other corpora are often more expensive.

2. The institution that created the annotation holds its copyright and can decide the distribution conditions. As modern corpora may comprise several annotation layers created by more than one research group, each group can be considered the creator of its annotation layer and can decide its terms of distribution (as a consequence, every annotation layer should potentially comprise a complete metadata record). Commercially available software tools that were used in the annotation process might restrict the terms of distribution of the resulting data set as well.
3. After DH 2007, a downloadable version of CorpusMasker will be available on our web site under an Open Source licence (<http://www.sfb441.uni-tuebingen.de/c2/>).
4. For centuries, typographers and graphic designers use the "Lorem ipsum dolor sit amet" text fragment to evaluate new layouts without resorting to writing actual text. The blind text gives the impression of a natural distribution of characters and whitespace without distracting the reader by conveying any meaning that could be interpreted intuitively. This approach might be useful for visualising masked corpora by means of XML to SVG transformations (Piez, 2004).
5. In a message posted to Corpora-List on Aug 19th, 2006, Péter Halácsy suggested an interesting method to distribute a copyrighted corpus under "fair use" conditions. Part of the copyright notice Halácsy et al. apply to the Creative Commons-based licence of the "Hunglish" corpus (D. Varga et al, 2005) reads: "We prevented the illegal use of copyrighted material by shuffling the texts at sentence level. This form is still useful for research purposes, while it does not infringe upon the rightholders' interests."

Literate Documentation for XML

Kevin M. Reiss *(kreiss@gc.cuny.edu)*
Mina Rees Library
Graduate and University Center
City University of New York

The Current State of XML Documentation

Current practices for the documentation of XML schema have not progressed much beyond the recommendations made in Eve Maler and Jeanne El Andaloussi's (1996) text Developing SGML DTD's: From Text to Model to Markup. That volume recommended an approach to DTD design called document type modeling to SGML DTD designers that, if followed, produces both a reusable and customizable DTD and a well-structured reference manual that documents that DTD.

Despite the development of XML itself, multiple XML schema languages, and the proliferation of XML applications across many disciplines this is still about all the documentation most XML applications provide today. In fact in many cases the documentation provided is much more inconsistent and haphazard than that recommended by Maler and El Andaloussi. The lack of consistent and structured prose documentation is especially problematic to humanists who use XML. To a humanist creating finding aids, marking up historic texts, or preparing a text for linguistic analysis the clear communication of the markup language designer's intentions is critical in making decisions on how to properly apply markup.

The problem of interpreting the designer's intentions is compounded by the fact that XML lacks a formalized, machine-readable knowledge representation technology that allows a designer to explicitly represent in unambiguous fashion the semantics of a XML application (Renear et al., 2002). Given the difficulties that researchers working in this area have made towards realizing a system that can reliably represent machine-readable markup semantics (Dubin et al., 2002)(Dubin et al., 2003) (Dubin, 2003) (Dubin and Birnbaum, 2004)(Marcoux, 2006) it may be more profitable to experiment with a system that can improve the quality of the prose documentation made available with XML applications.

This project proposes a XML schema authoring environment based on the literate programming paradigm that also contains constructs that force the markup language designer to

consistently and unambiguously document the application using natural language. This approach can help bridge the gap between current XML documentation practices until tools that can specify machine-readable semantics for XML appear in the future.

Literate Programming and XML

Literate programming directs the author specify both the program source and the prose documentation for that code within the same file (Knuth 1984). This file is then fed through a processor that produces both executable program code and formatted prose documentation for the program. A XML literate programming environment produces both a validating schema and prose documentation for that schema. The experimental XML literate programming system sketched out in this project utilizes the Text Encoding Initiative's One Document Does it All (ODD) literate programming system.

The TEI P5 (Sperberg-McQueen and Burnard, 2005), the most recent revision of the guidelines, has substantially updated the ODD system. The guidelines now use Relax NG for the validating schema component and formally include and document the elements and attributes that make up the ODD system (Burnard and Rahtz, 2004). This new module (Chapter 27) provides TEI users a formal way to document local customizations of the TEI and produce consistent, well-formatted prose documentation for these extensions. The documentation chapter of the TEI P5 describing the ODD states that the ODD is not restricted to just serving as the means to document and generate the TEI, but can also be used to document and produce a schema for any type of markup scheme (Sperberg-McQueen and Burnard, 2005).

The ODD+

This project takes advantage of the general purpose component of the ODD to create an experimental general purpose XML literate programming system that could be termed the ODD+. The support for modularity, element and attribute classes, schema customization, and XML namespaces within the ODD (Burnard and Rahtz, 2004) make it an ideal tool to make up the core of a literate programming system for XML. The advance of the ODD+ system will be the provision of a "semantic checklist" that the author must complete for each element and attribute in a XML application. This checklist will force the markup language designer to document new elements or attributes with precision and consistency. If the checklist is not properly completed the ODD+ processor will fail. The ODD+ application will be implemented by using the TEI P5 extension mechanism to extend the current ODD module to include the elements and attributes that will be necessary to implement the ODD+.

Identifying the Semantic Checklist

What sort of questions will be included in the semantic checklist? The researchers who have investigated the question of XML semantics have identified a number of constructs that occur within markup that seem appropriate candidates for inclusion in the semantic checklist (Renear et al., 2002)(Sperberg-McQueen, Huitfeldt, and Renear, 2000). The questions identified by these researchers include:

1. Can an element have different meanings depending on the content within it is used?
2. Issues of propagation: does the property declared by a given element or attribute apply to child elements, their attributes, and character data contained within them?
3. Issues of class memberships:
    1. Does the element or attribute serve as a superclass?
    2. Is the element or attribute derived from another?
4. Relationships to other elements:
    1. What are the properties that a parent child relationship implies?
    2. What are the properties that a sibling relationship implies?

The author will report on the difficulty of integrating the above questions and others into a workable semantic checklist that can be deployed within a general purpose literate programming system for XML like the ODD. The author will experiment with the ODD+ system and widely used XML applications in the humanities, such as the TEI itself or the Metadata Encoding and Transmission Standard (METS). The author will work with interesting subsets within each application and produce documentation for them using ODD+. This will illustrate the viability of the ODD+ as a potential general purpose literate programming tool for XML.

The author will experiment with different techniques that will strive to ensure consistency of format and language usage in the prose documentation produced by the ODD+. These may come in the form of Schematron rules written to ensure that the author uses language consistency when discusses issues of propagation for example. The author may also experiment with XSLT templates that may contain the outlines of statements that describe class relationships within a XML schema. These outline statements could then be filled in with the specific element or attribute names of the language designer's schema. This approach is an approximation of the skeleton sentence

technique suggested by Sperberg-McQueen, Huitfeldt, and Renear (2000). The author will report the results of these various experiments and present a demonstration of the ODD+ system.

Bibliography

Burnard, Lou, and Syd Bauman, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium, 2005. <http://www.tei-c.org.uk/P5/Guidelines/index.html>

Burnard, Lou, and Sebastian Rahtz. "RelaxNG with Son of ODD." *Proceedings of Extreme Markup Languages 2004, Montréal, Québec, August 2004.* Ed. B. T. Usdin. Montréal, Canada, 2004.

Dubin, David, and David J. Birnbaum. "Interpretation Beyond Markup." *Proceedings of Extreme Markup Languages 2004, Montréal, Québec, August 2004.* Ed. B. T. Usdin. Montréal, Canada, 2004.

Dubin, David, C. M. Sperberg-McQueen, Allen Renear, and Claus Huitfeldt. "A Logic Programming Environment for Document Semantics and Inference." Paper presented at ALLC/ACH 2002, Tübingen, Germany, July 2002. 2002.

Dubin, David, C. M. Sperberg-McQueen, Allen Renear, and Claus Huitfeldt. "A Logic Programming Environment for Document Semantics and Inference." *Literary & Linguistic Computing* 18.2 (2003): 225-233.

Dubin, David. "Object Mapping for Markup Semantics." *Proceedings of Extreme Markup Languages 2003, Montréal, Québec, August 2003.* Ed. B. T. Usdin. Montréal, Canada, 2003.

Knuth, Donald. "Literate Programming." *The Computer Journal* 27 (1984): 97-111.

Maler, Eve, and Jeanne El Andaloussi. *Developing SGML DTDs: From Text to Model to Markup.* Upper Saddle River, NJ: Prentice Hall PTR, 1996.

Marcoux, Yves. "A Natural-language Approach to Modeling: Why is some XML so Difficult to Write?" *Proceedings of Extreme Markup Languages 2006, Montréal, Québec, August 2006.* Ed. B. T. Usdin. Montréal, Canada, 2006.

Renear, Allen H., David Dubin, C. M.Sperberg-McQueen, and Claus Huitfeldt. "Towards a Semantics for XML Markup." *Proceedings of the 2002 ACM Symposium on Document Engineering, McLean, VA, November 2002.* Ed. R. Furuta, J. I. Malectic and E. Munson. New York: Association for Computing Machinery, 2002. 119-126.

Sperberg-McQueen, C. M., David Dubin, Claus Huitfeldt, and Allen H. Renear. "Drawing Inferences on the Basis of Markup." *Proceedings of Extreme Markup Languages 2002, Montréal, Québec, August 2002.* Ed. B. T. Usdin and S. R. Newcomb. Montréal, Canada, 2002.

Sperberg-McQueen, C. M., Claus Huitfeldt, and Allen H. Renear. "Meaning and Interpretation of Markup." *Markup Languages: Theory and Practice* 2.3 (2000): 215-234.

Digital Text Projects in Eastern Europe: Promoting International Interoperability

Miranda Remnek *(mremnek@uiuc.edu)*
Slavic & Eastern European Library
University of Illinois at Urbana-Champaign

The widely divergent languages and cultures of Russia, Eastern Europe & Eurasia (hereafter Eastern Europe), present a rich terrain for digital scholarship, but owing to a number of factors—including poorly-endowed or mutually-incompatible infrastructures, language barriers, and different approaches to the construction of metadata—the corpora already in existence and those in development are often poorly known elsewhere, and in many cases are hidden from international resource discovery and data sharing.

In view of this situation, it was clear to the Digital Projects Subcommittee of the *American Association for the Advancement of Slavic Studies* (<http://www.aaass.org>) that improved documentation would constitute an important step toward the promotion of knowledge sharing and interoperability.[1] The Slavic & East European Library at UIUC is therefore building an international *Inventory of Slavic, East European and Eurasian Digital Projects* (<http://www.library.uiuc.edu/spx/inventory>). The *Inventory* already describes at least 360 collections in 110 projects, and we also foresee web submissions from international partners to promote collaborative registration. But even now the *Inventory* is heavily used; recent statistics indicate that in the period July 2005-June 2006 it received around 25,000 hits. It has also become an OAI Data Provider, and although inventory contents (unlike the digital collections they record) are sometimes not reflected in aggregated search services like OAIster (<http://oaister.umdl.umich.edu/o/oaister/>), the *Inventory's* records have indeed been harvested and now display in OAIster searches—likely to improve dramatically the visibility of substantive digital projects in the Slavic field.

But the project goals go far beyond the compilation of a mere inventory. Besides the implementation of intensive content development, promotion of metadata standardization, expanded reference assistance and interactive user options, the Inventory will focus on an expansion of the OAI data provider system and item-level harvesting. (It also hopes to explore customized delivery services and an archival framework for at-risk collections in the Slavic field, currently being explored through an inter-institutional grant proposal).

Nevertheless, the project team acknowledges that metadata practices in Eastern Europe are still very diverse, and far from standardized. At one end of the scale, where relatively simple encoding is involved, even schemes like Dublin Core are poorly recognized. This makes the implementation of relatively well-established data-sharing protocols like OAI-PMH (*Open Archives Initiative: Protocol for Metadata Harvesting*) little known in Eastern Europe. A registry of OAI data providers at UIUC demonstrates that while OAI data providers are well established in the US and Western Europe, they are much less widely encountered further East.[2] True, the implementation of OAI data interchange is not necessarily well standardized in countries with far more providers.[3] Furthermore, the whole question of the efficacy of OAI-PMH as a metadata transfer protocol is still somewhat open to question.[4] Nevertheless, its power is generally well-recognized, and its further penetration into Eastern Europe is certainly a desirable goal.

At the other end of the scale, the level of deep encoding is even less widely standardized in Eastern Europe. The decade-long Institute of World Literature project in Moscow known as FEB (*Fundamental'naia elektronaia biblioteka*) (<http://www.feb-web.ru/>) has an enviable collection of over 50,000 Russian literature texts, all heavily encoded in SGML—but not according to the *Text Encoding Initiative* (TEI) Guidelines. True, the TEI is not a standard, but interoperability with such an impressive collection would be highly desirable. Other such examples are available. To be sure, there are exceptions: the recent TEI meeting in Sofia, Bulgaria revealed the extent to which TEI projects are beginning to gain root in South Eastern Europe.[5] There are also similar developments in a more abstract sense. The TEI Guidelines have been translated into Russian,[6] and efforts are also being made to promote the further internationalization of the Guidelines, and to ensure their availability in a greater number of languages, including Bulgarian.

Yet even when East European scholarly projects adopt the TEI, standardized approaches to the conversion of metadata to OAI-compliant formats are also at issue, and there is evidently much ground still to be covered. Hence, the object of this paper will be to:

1. Design and conduct a survey of metadata practices in a number of East European digital centers (including existing UIUC partners) that are sponsored both by institutions and by individual faculty teams;
2. Produce an up-to-date analysis of their awareness, evaluation and observance of international metadata standards;

3. Identify problems and practices preventing their implementation;
4. Suggest ways in which steps can be taken to ameliorate this situation, including crosswalks and other technical procedures.

The results of this exercise will not only assist in a practical sense the development of the *Inventory of Slavic, East European and Eurasian Digital Projects* (the only registry of substantive, scholarly East European digital initiatives), but will also result in better awareness and information-sharing among Slavic digital practitioners. This is especially vital as the institutional repository movement gains ground.[7] Experience shows, particularly in West Europe, that the issue of standardized metadata is paramount, and that subject repositories are becoming of even greater interest since they tend to be populated by scholars working on their own projects and producing more substantive metadata. In this environment, East European scholars should not excluded from the pool of informed experience.

---

1. Other points in the group's charge include: (2) fostering informed participation in future initiatives; (3-4) increasing digital information and training opportunities for Slavic librarians; (5) complementing national efforts to establish digital repositories. See <http://www.library.uiuc.edu/spx/BnD/DigPro.htm#charge>
2. See <http://gita.grainger.uiuc.edu/registry/ListTLDs.asp>. U.S. providers are proliferating: in the past year the number of "edu" domains increased from 265 to 332, and "org" domains from 149 to 182. Likewise in Western Europe: U.K. DPs went from 89 to 114, Germany from 73 to 99. But in Eastern Europe it is different; the totals for Hungary (4), Czech Republic (2) and Slovenia (2) stayed the same this year, and Russia increased only slightly, from 3 to 4. Only Poland increased significantly, from 4 to 18.
3. As reported at the Oct. 2005 OAI workshop in Geneva, the German Initiative for Networked Information (DINI) has installed a certificate system in order to bring OAI providers into greater standardization.
4. The Open Archives Initiative OAI-Implementers group announced on Nov. 3 2005 that it is "studying the effectiveness of OAI and some other related methods of creating interoperable online libraries" and has "posted a questionnaire 'Survey on Common Interface Frameworks for Online Libraries' on the web."
5. See Milena Dobreva, "TEI in South-Eastern Europe: Experience and Prospects," TEI Members Meeting, Sofia, Bulgaria, October 28, 2005.
6. See <http://www.tei-c.org.uk/Lite/teiu5_ru.rtf>.
7. Here again the number of East European repositories lags far behind. In July 2006 there were 144 DSpace instances worldwide, and less than half were in the U.S. and U.K. (62), indicating that many non-English-speaking countries (like Japan) are aware of and attracted to this increasingly popular open source software. But in Eastern Europe only 2 DSpace sites were listed (in Russia). This lack of awareness was substantiated at an e-text conference in Eastern Russia in July 2006 entitled *Modern Informational Technologies and Written Heritage: From Ancient Manuscripts to Electronic Texts* at which I gave a presentation entitled "Promoting the TEI in scholarly communities: OAI harvesting, digital repositories." Conference attendees described many sophisticated projects, but there was little familiarity with OAI or digital repository software used in the West. One scholar was familiar with these technologies, but in a presentation entitled: "A Computer System for the Creation and Maintenance of Electronic Collections of Ancient Texts," V. S. Iuzhikov (Kazan State University) wrote: "For the creation and development of electronic libraries there are several systems. Among the best known are Greenstone and DSpace. But they are oriented mostly toward text libraries with small number of illustrations, which are poorly suited for the collection of old printed works. Therefore it is less time-consuming and gives better result to build a local system." Given the specialized nature of his material, this conclusion was understandable. Yet the general lack of East European involvement with these international standards and methodologies remains true.

Modeling, Explanation, and Ontology in the Cultural Sciences

Allen H. Renear *(renear@uiuc.edu)*
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign

Abstract

Humanities computing must make *conceptual modeling* one of its defining activities. This is necessary not only to provide more effective computational support for scholarship in the humanities, but also in order for humanities computing itself to become something more than a "bag of tricks". This paper describes promising developments already underway in several areas within humanities computing, generalizes from past, if partial, successes (text encoding), draws parallels with recent work in bioinformatics, and argues that the logical identity of conceptual modeling and theory-construction makes conceptual modeling simply a computationally oriented variant of scientific explanation, and, more broadly, interpretation and understanding in general. As an example of how model development can provide fundamental insights into the cultural world a case is described where a set of ontology evaluation criteria developed within computer science is applied to an influential framework for cultural entities and the result that a radical reconfiguration of that framework is suggested.

Acknowledgements: There is undoubtedly much here that is influenced by seminal papers on these issues by Willard McCarty and John Unsworth; I acknowledge a debt to them without of course any suggestion that they see things as described here.

Introduction

The nature of humanities computing has long been problematic in the digital humanities community. Part of the reason for this anxiety is that humanities computing is struggling to move out of a pre-scientific phase. "Science" here is of course intended to be taken in the broadest possible sense — the point is not to separate disciplines such as physics on the one hand from, say, literary studies on the other, but rather to distinguish practices that have primarily non-epistemic non-theoretical objectives from those that have understanding, explanation, and knowledge as their principal immediate goals. Nor is there any implication of a clear boundary between scientific practices and non-scientific practices, or that there isn't a complex and essential mutual involvement — the claim is only that some rough distinction of this sort is possible and useful.

If humanities computing is to mature as a coherent field of intellectual inquiry it will need to make a transition from being an *ad hoc* set of useful techniques, to more systematic general methods and theories that make direct contact with the specific methods and theories of the disciplines it supports. This theme is not new, but our progress has been modest so far and some further discussion and elaboration is in order.

First, there are new reasons to be optimistic. For one thing this progress is already occurring in other informatics fields—bioinformatics as an example—and the specifics of these changes provide an indication of how things might go in humanities computing as well. Moreover the growing significance of conceptual models and ontologies that is already in evidence in humanities computing, and the nature of the discourse around them, is a sign that this transition is perhaps already underway. Finally one of the most theoretically productive areas within the humanities computing community, namely text encoding, arguably has been as successful as it has precisely to the extent that it has embodied the features recommended here — and owes its limitations and disappointments to the extent to which it has not. What all these things have in common is the foregrounding of conceptual modeling, explicitly or implicitly.

I use the term "conceptual modeling" broadly, meaning any formally defined abstract representation of a domain of interest. Typical examples of modeling languages are the ER and UML diagrams used in business applications, logic-based knowledge representation formalisms such as description logics or the frame languages (such as the KL-ONE family) common in artificial intelligence, and the various formal ontologies now used in computational biology and elsewhere

Science in the broad sense is the development of *understanding*. Scientific understanding comes typically in the form of theories, and theories are at least in part the systematic identification of objects and relationships that are the salient features of some domain. Informatic disciplines must put these theories, formalized as conceptual models that support computational processing, at the center of their identity. And for informatic disciplines to be sciences themselves they must not only exploit such conceptual models, but participate in their development and exploration. This is true for informatics in general, and for humanities computing in particular.

Modeling in Bioinformatics

Other areas that began as craft-like bag-of-tricks approaches to applying computing to a scientific field have now begun to evolve a more coherent body of method, theory, and metatheory. Perhaps the most dramatic case is bioinformatics. Although initially a haphazard and opportunistic application of computing techniques, there are now a considerable number of families of well-theorized methods tightly tied to biological theory. Of particular interest here is the role of so-called *ontologies* in contemporary bioinformatics such as the Gene Ontology (Ashburner et al., 2000), the Foundational Model of Anatomy (Rossee & Mejino, 2003), and other ontologies in genomics, molecular function, neuroscience, biodiversity, and other areas of the life sciences as well. An enormous amount of effort is going into the development and use of these formal models.

In some cases the influence on biological science *per se* has still been modest, but generally the results are extremely promising and in a few cases the influence has been stunning. The Gene Ontology in particular, has been enormously successful and provides the overall framework for contemporary genomics. Of special significance for us is the extent to which the fundamental explanatory entities and properties of biological theories are explicitly identified as the core constituents of these conceptual models, blending the work of first order science with its bioinformatics support.

Modeling in the Cultural Sciences

Within the humanities computing community text encoding systems such as the TEI can be seen as a step in the same direction as the new ontology-based models in computational biology. Behind the methods of the TEI is at least arguably an informal positing of text models, components, and relationships. However unlike the bioinformatic ontologies the models being indicated by text encoding techniques are only indirectly and implicitly identified. This has been pointed out by a number of critics (Raymond & Tompa, 1992; Raymond, Tompa, & Wood, 1996, 1998, 2001; Buzzetti, 2002), and projects to develop remedies are underway (Sperberg-McQueen et al., 2002; Renear et al., 2003).

Another promising sign are the ontologies being developed in museum studies CIDOC/CRM (Crofts et al., 2003) and library science (IFLA 1998). For the most part these are being put forward as frameworks for designing digital information management systems or shaping the development of standards, policies, and procedures, but to the extent that they succeed at efficient, functional, and interoperable systems they may be considered confirmed as substantive theoretical proposals for how to understand the nature of the cultural world. And where these models and ontologies appear to be applications of existing explanatory theories of cultural objects and relationships that is more evidence that they are achievements which advance our understanding of that world.

An Example: Revising FRBR

What follows is an example intended to support two claims. First that the evolution of conceptual models intended to support information systems design can in fact be an activity of first order science (in the broad sense), and second that precisely the same techniques designed for general evaluation of ontologies in natural science can be effectively employed in evaluating cultural ontologies.

Several models for cultural objects identify such things as works, texts, editions, and individual items as *types* of things, fundamental kinds. Not only are various sorts of conceptual arguments advanced in favor of such an analysis, but so is the relative success of such frameworks in guiding the design of effective information systems that are effective, efficient, and interoperable. That is, the conceptual and practical success of these systems can be taken as indicating that they correctly represent *how things are*. A well-known model of this sort is IFLA's Functional Requirements for Bibliographic Records (IFLA 1998), a framework designed primarily to support library cataloguing, but which clearly has wider application — FRBR in fact describes itself as a "conceptual model of the bibliographic universe".

A puzzle recently emerged about an assignment of one particular cultural object (XML documents, as defined in the W3C XML standard) to the appropriate FRBR entity type. Some considerations argued for an assignment to the FRBR *expression* (roughly: text) entity type, and other competing considerations argued for assignment to the FRBR *manifestation* (roughly: edition, or format) entity type. The proposed resolution was an awkward and unsatisfying exception to the FRBR framework which preserved the ontology of works, expressions, editions, at the cost of complexities elsewhere. (Renear et al 2002, Renear 2005).

Two years later however it was noted that when a proposed criterion for ontology evaluation which had been developed for scientific and general purpose ontologies (Guarino & Welty, 2000; Guarino & Welty, 2002) was applied to the FRBR framework the result was an anomaly that suggested refactoring the four Group 1 entity types into new entities and corresponding roles. (Renear, 2006). Under the resulting revised model the manifestation and expression entity types are not treated as true entity types (fundamental kinds), but rather *roles* (relationships) that entities of other types have in particular

circumstances—and in fact this revision was extended to all four FRBR entities (Renear & Dubin, forthcoming). We realized that this new way of conceptualizing cultural objects appeared to be an application of a more general approach developed by John Searle (1995) as well as consistent with some work in aesthetics (Levinson 1980). In short, all four FRBR "entities" can be reinterpreted not as entities but as roles that other non-cultural things have under specific social contexts of "collective intentionality".

That this is not just a practical adjustment in a conceptual model for humanities *computing*, but a substantive claim in the science of cultural objects in general, can be established by comparing it to familiar claims about cultural objects by scholars such as Ingarden, Wellek, Richards, Fish, Goodman, Tanselle, Schillingsburg and others. In this comparison contradiction will serve as well as convergence to make the point that these assertions fall within the domain of humanist inquiry, and not merely within some auxiliary practice.

The Relative Neutrality of these Recommendations

It may be thought that putting conceptual modeling at the heart of humanities computing and seeing the development of models as first order scholarship in the humanities requires accepting some archaic and dubious philosophical view, a positivistic scientism perhaps, or even philosophical realism. But this is not so. It is possible to engage in ontology without taking a philosophical *meta*-ontological position. Many different philosophical positions (including constructivism and relativism) are all perfectly consistent with taking conceptual modeling as the natural activity of humanities computing, and at least part of humanities scholarship more generally. Of course the *results* of modeling will not be neutral vis-à-vis other first order theories, but the *activity* of modeling need not be exceptionally controversial. Again the experience of bioinformatics, where researchers with different philosophical views (or no philosophical views at all) nevertheless often agree, or more importantly disagree, on specific issues without rejecting the overall approach, is significant.

This is not to say that the position is entirely neutral. It may well conflict with some general accounts of humanistic inquiry which hold that such inquiry is radically different than the natural sciences, especially if those accounts are understood as applying in general or without exception to the entire range of humanistic scholarship — although I am not yet convinced that it does in fact conflict with the traditions of hermeneutics and *verstehen*, despite the differences in tone and direction. In any case that the recommendation is not entirely uncontroversial cannot be taken as a decisive mark against it at the outset. Whether it is sound or not will depend on how it fares against its competitors in improving our knowledge and understanding of the cultural world.

Concluding Remarks

I have argued that humanities computing must make conceptual modeling its central defining activity if it is to fully realize its promise to substantially advance humanities scholarship, and that such modeling is an intrinsic part of first order humanities inquiry, and not merely an auxiliary activity.

Bibliography

Ashburner, Michael, et al. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25 (2000): 25-29.

Buzzetti, Dino. "Digital Representation and the Text Model." *New Literary History* 33.1 (2002): 61-88.

Cover, Robin. "XML and Semantic Transparency." 1998. <http://xml.coverpages.org/xmlAndSemantics.html>

Cover, Robin. "Conceptual Modeling and Markup Languages." 2001. <http://xml.coverpages.org/conceptualModeling.html>

Crofts, Nick, I. Dionissiadou, Martin Doerr, and Matthew Stiff. *Definition of the CIDOC Object-Oriented Conceptual Reference Model, ISO/TC46/SC4/WG9/N2, International Organization for Standardization.* 2003.

Guarino, Nicola, and Christopher Welty. "Evaluating Ontological Decisions with OntoClean." *Communications of the ACM* 45.2 (2002): 61-65.

Guarino, Nicola, and Christopher A. Welty. "A Formal Ontology of Properties." *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management Lecture Notes In Computer Science* 1937 (2000): 97-112.

International Federation of Library Associations. Ed. Marie-France Plassard. *Functional Requirements for Bibliographic Records: Final Report.* UBCIM Publications-New Series. München: K.G.Saur, 1998.

Levinson, Jerrold. "What a Musical Work Is." *The Journal of Philosophy* 77.1 (1980): 5-28.

McCarty, Willard. "Humanities Computing: Essential Problems, Experimental Practice." *Literary and Linguistic Computing* 17.1 (2002): 103-125. doi:10.1093/llc/17.1.103

McCarty, Willard. "Modeling: A Study in Words and Meanings." *A Companion to the Digital Humanities*. Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing Ltd, 2004. 254-70.

Raymond, Darrell R., Frank Wm. Tompa, and Derick Wood. *Markup Reconsidered*. Department of Computer Science, Technical Report No. 356. The University of Western Ontario, 1993. Presented at the First International Workshop on the Principles of Document Processing, Washinton DC, October 21-23 1992; an earlier version was circulated privately as "Markup Considered Harmful" in the late 1980s.

Raymond, Darrell R., Frank Wm. Tompa, and Derick Wood. "From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML." *Computer Standards & Interfaces* 18.1 (January 1996): 25-36.

Renear, Allen H. "Is An XML document a FRBR Manifestation or a FRBR Expression? — Both, Because FRBR Entities are not Types, but Roles." *Proceedings of Extreme Markup Languages, Montréal, Québec, August 2006*. 2006. <http://www.idealliance.org/papers/extreme/Proceedings/html/2006/Renear01/EML2006Renear01.html>

Renear, Allen H., David Dubin, C. M. Sperberg-McQueen, and Claus Huitfeldt. "Towards a Semantics for XML Markup." *Proceedings of the 2002 ACM Symposium on Document Engineering. McClean, Virginia, November 2002*. Association for Computing Machinery, 2002.

Renear, Allen H., and David Dubin. "The Four FRBR Group 1 Entities are Roles, Not Types." Forthcoming.

Renear, Allen H. "Text from Several Different Perspectives, the Role of Context in Markup Semantics." *Atti della conferenza internazionale CLiP 2003, Computer Literacy and Philology, Firenze, 4–5 December 2003*. Ed. C. Nicolas and M. Moneglia. Florence: University of Florence Press, 2005. < http://people.lis.uiuc.edu/~renear/clip2005.pdf>

Renear, Allen H., Christopher Phillippe, Pat Lawton, and David Dubin. "An XML Document Corresponds to which FRBR Group 1 Entity?" *Proceedings of Extreme Markup Languages, Montréal, Québec, August 2003*. Ed. B. T. Usdin. 2003.

Rosse, Cornelius, and José L. V. Mejino, Jr. " A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy." *Journal of Biomedical Informatics* 36.6 (2003): 478-500.

Searle, John R. *The Construction of Social Reality*. New York: The Free Press, 1995.

Sperberg-McQueen, C. M., David Dubin, Claus Huitfeldt, and Allen H. Renear. "Drawing Inferences on the Basis of Markup." *Proceedings of Extreme Markup Languages 2002, Montréal, Québec, August 2002*. Ed. B. T. Usdin and S. R. Newcomb. 2002.

Sperberg-McQueen, C. M., Claus Huitfeldt, and Allen H. Renear. "Meaning and Interpretation of Markup." *Proceedings of Extreme Markup Languages, Montréal, Québec, August 2000*. Graphic Communications Association, 2000.

Unsworth, John. "Scholarly Primitives: What Methods do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" Paper presented at Symposium on Humanities Computing: Formal Methods, Experimental Practice. King's College, London, May 13, 2000. 2000. <http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html>

Unsworth, John. "What is Humanities Computing and What is Not?" Distinguished Speakers Series, Maryland Institute for Technology in the Humanities, University of Maryland, College Park MD, October 5, 2000. 2000. <http://www.iath.virginia.edu/~jmu2m/mith.00.html>

Unsworth, John. "New Methods for Humanities Research." The Lyman Award Lecture, National Humanities Center, November 11, 2005. 2005. <http://www3.isrl.uiuc.edu/~unsworth/lyman.htm>

Unsworth, John. "Knowledge Representation in Humanities Computing." n.d.. <http://www.iath.virginia.edu/~jmu2m/KR/>

The AXE Tool Suite: Tagging Across Time and Space

Doug Reside *(dreside@umd.edu)*

Maryland Institute of Technology in the Humanities (MITH)

Since the codification of the Text Encoding Initiative standards in the mid-1990s, the process of the creation of digital editions and archives is largely one of "marking up" existing texts in XML. Originally, this was often done "by hand." Scholars would add XML tags to existing text documents using a text editor-a tedious process often fraught with errors. Over the last decade several tools have been produced which make this process somewhat more efficient and accurate, though most still require more than a beginners familiarity with XML encoding, and few are open source. Moreover, many digital humanities projects have, of late, become far more multi-medial--relying on image, video, and audio files as well as text. Existing markup tools have only begun to work with these non-textual artifacts. As digital archives continue to grow, the markup tools used to encode them must become more flexible and easier to use. The Ajax XML Encoder (AXE), developed at the Maryland Institute for Technology in the Humanities (MITH) is intended to be the tool suite that meets this need.

AXE is a free tool suite that will allow users with limited technical skills to deeply tag text, images, video, and audio files for inclusion in digital archives. The program combines and extends the functionality of proprietary online tools such as YouTube and Flickr in a (mostly) open source[1], web-based platform for scholarly use. The tagging tool is designed in AJAX and Macromedia Flash and generates and uses a MySQL database. Like the mythological Ajax's axe, MITH's AXE provides users with enough power and flexibility to accomplish their tasks without a great deal of assistance from higher powers (the technical or professorial "gods" graduate student workers must often invoke when using earlier encoding tools).

Users of this tool are divided into two classes: managing editors and editor-users. Managing editors are permitted to define the sorts of tags all users can use and are able to remove items from the database. All users, however, are able to add new multimedia content to the database, tag it, and search for preexisting content in the database. Tagging is naturally accomplished in different ways depending on the media. For sound files, an mp3 file is played through a Flash plug-in. When the user pauses a sound file once, the timing is recorded. When the user pauses the file again, the end time is recorded and the user is presented with a web-based form to tag the selection. If appropriate, a textual transcript of the sound with corresponding times can be made. Providing a tool for image tagging was more difficult. Although popular websites like Myspace and Facebook provide some models for folksonomic image tagging, the tag-spaces are usually limited to regular polygons (like squares) and do not store the resulting data in a form that can be easily shared. A better model for image tagging is the ARCHway Project's "Image Tagger," described in the June 2006 issue of the International Journal on Digital Libraries by designers Dekhtyar, Iacob, Jaromczyk, Kiernan, Moore, and Porter . The program is hindered, though, by the clumsy Java-based interface that requires users to install and learn specialized software. Our program uses an AJAX website which allows the user to add points to an image map (represented on the image itself via a 1 pixel `div` element with a red border) [see figure 1]. Once the image map is drawn, the user can describe it with tags defined by the editor (or editors) of the project. The entire image can also be so described

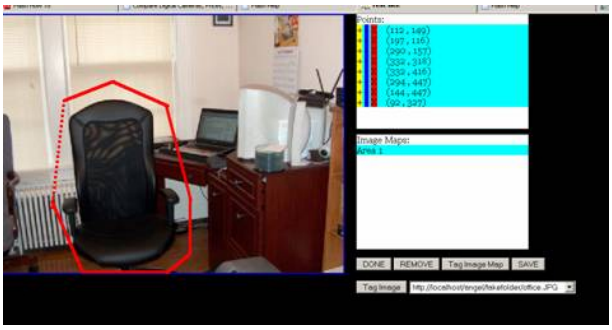


Figure 1

Video tagging, as one might expect, uses a combination of the image and audio tagging methods. The video is first converted to flash movie format using FFMpeg. The audio portion of the movie is handled in exactly the same way as stand alone audio. The visual portion of the movie is handled with a combination of the techniques used for images and sound. The user can tag, for instance, the start and end times of a particular segment and add metadata to this portion. AXE even allows users to tag images within the frames of movie files. If the user wished, for instance, to tag a tree in the background of a shot, the user first marks the time period in which the appropriate image appears. A user-defined number of the frames from this segment are then stacked and rendered translucent. With some adjustments, the user can then click a series of coordinates which define the selection over the space of the segment. This process is rendered through Javascript and Macromedia's Flash.

Once all the tags have been created, they are parsed into a mySQL database (the XML is generated first to allow users to work offline and then feed the XML to the database later). This

database can be searched through traditional keyword or tag cloud searches, but the interface also allows guided browsing. When the user views a multimedia object, the user is also presented with a series of "related" objects (after the e-commerce model in which customers are presented with a series of products related to one they are currently viewing). If the user selects a sub-element in the document (perhaps something like a tree in the background of a photograph), the "related" documents will change to reflect relationships centered on the user's choice. The user can also set global limits to what suggestions are made. The related documents will change to reflect user selection of sub-elements. The user can set global limits on what relations are presented (e.g. related documents must be dated after the current document). New tags and documents from remote sites can also be added to the database by editor-users. Later users can decide whose tags they trust and block those by anonymous or unreliable taggers.

AXE is currently being used in a project headed by Angel David Nieves to create a multi-perspective history of a 1976 student uprising in Soweto, South Africa. Dr. Nieves has an enormous wealth of multimedia primary documents surrounding the event and hoped that the mass of materials would provide new ways of telling the story. We needed, however, to process and present this material in an unbiased way that allowed for multiple interpretations of the events. AXE is being used to create an interface which, as much as possible, left the arrangement and interpretation of the materials up to the individual users.

AXE may eventually prove essential in the creation of digital library archives. Although academic libraries are now fairly adept at the digital preservation of textual material, few libraries provide searchable digital archives of sound and video. As the average available bandwidth of both users and institutions increase, software will be needed to allow the cataloging and access of this material. AXE seeks to fill this need.

1. Although all code written at MITH will be published on the web, the use of Adobe's Flash prevents the project from being truly open source.

Digital Humanties! The Musical

Doug Reside *(dreside@umd.edu)*

Maryland Institute for Technology in the Humanities (MITH)

It has now been twenty-five years since the development of the first personal computer, and while many humanities scholars have begun to explore how these machines might assist them in their work, there have been very few attempts to consider how electronic tools might be used to study musical theater. Musical theater seems particularly well-suited, though, to the multimedia capabilities of the modern PC. David Saltz's *Virtual Vaudeville*and the "Web Operas" on Paul Howarth and Jim Farron's *Gilbert and Sullivan Archive* prove that dramatic text and music can be presented together in interesting ways through a digital interface. However, neither is precisely the right model for presenting musical theatre texts. This presentation will describe the particular suitability of contemporary musical theater for electronic presentation and will demonstrate the author's electronic edition of the musical *Parade.*

The electronic edition of *Parade* serves as a prototype for an upcoming series of electronic critical editions musical theater texts the author is currently working with the Library of Congress to develop. Using the standards designed by the Text Encoding Initiative (TEI), the text of the Broadway and the touring version of the libretto of *Parade* were encoded into one file. Textual, critical, and explanatory notes were also encoded in the text using the same system. The text is presented in an interactive AJAX website. Users can choose how much of these texts to view and the way in which they should be displayed. This functionality allows users to compare varying versions of the text and score side by side, or (using transparency) even one on top of the other. When the user clicks the lyrics of a song, the music begins playing at the beginning of the associated section and stops at the end (or when paused by the user). Analytical tools such as an automated collation feature, an exhaustive concordance of every word in the libretto, and a tool which extracts all the lines for a particular character along with cue lines to aid actors in memorization of their parts are also included. Future editions will include tools that will allow users to analyze patterns in the textual, musical, and terpsichorean languages of the poem. For example, a concordance of words will be linked to a concordance of musical phrases and both, in turn, will be linked to a concordance of choreographed movements. Videos of dance numbers will be linked to computer animations of the

movements of individual dancers. The early versions of these features will also be demonstrated in this presentation.

This presentation will also describe how such editions will benefit musical theater studies. Musical theater scholarship is growing rapidly. An increasing number of scholarly works are being published by academic presses, and in the past year alone there were over fifty dissertations written on the topic. While this scholarship is promising, the academy as a whole remains relatively uninterested in the musical. A recent panel on composer-lyricist Stephen Sondheim was scheduled for the Modern Language Association's annual conference but was later cancelled for lack of interest. The musical is still usually relegated to popular culture studies and is rarely included in the syllabi of contemporary drama courses. In part, this may be a reaction to the preponderance of mostly formulaic works in the genre. Today, aside from revivals and the occasional limited run at Lincoln Center, artistic (as opposed to purely commercially driven) musicals can rarely generate enough investor interest to go on Broadway. The creators of artistic musical theater bemoan the fact there is not a venue for their work. In a recent article for Opera News, musical theater composer Michael John LaChiusa (composer-lyricist of many recent artistic musicals), writes of recent Broadway hits, "All sense of invention and craft is abandoned in favor of delivering what the artist thinks a musical should deliver."[1] Composer-lyricist Jason Robert Brown echoes this complaint in an interview with the author, "Only in [today's Broadway] musicals do things exist with the sole purpose of entertaining their people and not hoping to make them think about the world." It could be that these are nothing but the complaints of artists unable to achieve widespread commercial success, yet the fact remains that the Broadway musical has, in the last fifteen years, almost completely consisted of light comedy and pastiche.

Significantly, although non-musical plays suffer many of the same pressures, a few serious plays regularly open and close on Broadway every season. A play can afford to fail, not only because production costs for plays are usually lower than for musicals, but because it is viewed as "high art." Producers are more willing to lose money on plays because they are contributing to works of "cultural importance." The musical, on the other hand, is trapped in a degrading cycle. The scholarly community, which might elevate the cultural respectability of the musical, generally ignores the form because so few musicals are artistically interesting (thereby withholding the cultural blessing which might inspire producers to fund more interesting musicals).

One way to break the cycle may be to provide scholars with better access to the best works of musical theater. There is evidence that increased access to the material can lead to increased cultural respect. The Gershwin's *Of Thee I Sing*, the first musical for which the libretto was commercially published, went on to become the first musical to win Pulitzer prize. Unfortunately, the full text of a musical is often available only to theater practitioners who rent the performance texts from theater licensing companies for a production. Despite Stephen Sondheim's reputation, there is, at the time of this writing, no commercially published libretto of his *Merrily We Roll Along* or *Saturday Night* If Arthur Miller's or Tom Stoppard's work were available only under such conditions one wonders if the literary community would have been so willing to embrace them.

Of course, the extremely multi-medial nature of musical theater makes the traditional codex a clumsy tool for presenting editions of the texts. Further, musicals often exist in many different versions with no one text definitely representing the title. The musical *Show Boat* , for instance, may be better thought of as the sum of the total of all its major incarnations rather than as any one production. Although this notion of textuality is difficult to represent in print, it is well-suited for an electronic edition. Further, electronic editions will assist not only individual scholars of musical theater, but instructors of the form as well. Teaching close reading of musical theater texts has always been an awkward matter. The problem of linking text to sound and the limited available of important texts often force instructors to assign videotaped or movie versions of the works. This approach does not encourage careful study of the written words of the musical. If scholarly electronic editions of musical theater are eventually produced, it may be that musical theater will finally find a firm footing in academic scholarship. The first printed libretto of a musical won a Pulitzer Prize. It seems possible electronic editions will increase the respectability of an art form that is even now slowly finding its deserved place in literary scholarship. Perhaps scholarly interest will generate more funding for serious musicals. Optimistic thoughts, to be sure—perhaps better suited for a Rodgers and Hammerstein song than a conference presentation. Still, at the very least, electronic editions of musical theater will provide a new way of studying and experiencing this important art form.

1. Michael John LaChiusa, "The Great Gray Way," *Opera News* (August 2005): 30-35.

Why Take Games Seriously? Digital Humanities and the Study of Games

Jason C. Rhody *(jasonrhody@gmail.com)*
University of Maryland at College Park

Until recently, the study of games from a humanities perspective has been a surprisingly impoverished field, with the relatively few titles generally philosophical or ethnographic in nature. Yet the recent rise of game studies has generated a wealth of scholarship as interdisciplinary as the games themselves. What is it about computer games that caused such an increase in attention? Why do these games merit such scrutiny? With a robust industry comprised of both popular and independent studios that have produced thousands of virtual worlds and environments, computer games have become one of the fastest rising modes of new media expression. As Janet Murray asserts, "the largest commercial success and the greatest creative effort in digital narrative have so far been in the area of computer games" (*Hamlet on the Holodeck* 51). If the computer represents a "single new medium of representation, the digital medium, formed by the braided interplay of technical invention and cultural expression," (Murray, *New Media Theory Reader* 3), the game is its most prominent genre, simulating sport, adventure, exploration, war, economies, and even life itself. In fewer than fifty years, computer games have grown from allowing text-based adventurers to crawl through fictionalized caves to generating miles upon miles of virtual landscape inhabited by its digital citizens and with economies rivaling that of several real-world countries. Whereas a single white dot once floated across a dark screen in an abstraction of table tennis, players can now top-spin their way through the rankings at a virtual Grand Slam tournament.

This paper examines both the long and short history of gaming and its scholarship within a humanities context, arguing that games play a crucial role in any serious academic study of new media and within the overarching framework of the digital humanities. After an overview detailing how games have influenced and engaged with traditional humanities topics, I discuss three ways that games should be taken seriously in the digital humanities: as models for learning and rhetoric (as with the "Serious Games" movement); as objects that are part of the larger digital media ecology desperately in need of preservation; and as objects of critical study. In short, this paper addresses why, and how, humanities scholars should take games seriously.

I begin with a brief review of games' rich 'long history' in the humanities, which includes art both *about* games, such as the 11th-century Chinese painting of "Ladies Playing Double Sixes," and the artistry *of* games, such as the pieces from a 19th-century Indian chess set, whose delicate nature suggests that they are as much decorative as functional. From a cultural historical perspective, games reveal much about social hierarchies: the extravagance of the late sixteenth-century Moghul ruler Akbar, who had elaborate Pachisi boards constructed of flagstones so that he might play, using beautiful girls dressed in appropriate colors; or the inclusion of a game within the Chinese literati's four key accomplishments in the arts: mastery of the zither, calligraphy, painting, and weiqi (perhaps better recognized by its Japanese name: Go); or the simple modesty of card and board games found at all levels of society. From a literary perspective, games inflect upon the playfulness of Dadaism, the linguistic games of the Oulipo, or even as plot device, as with Scrooge, so taken with the party's games "wholly forgetting in the interest he had in what was going on, that his voice made no sound in their ears, he sometimes came out with his guess quite loud, and very often guessed quite right, too" (Dickens, *A Christmas Carol* 69).

The long history of games conspires with the shorter history of the modern computer game, which both remediates (in Bolter and Grusin's sense) previous artistic and ludic forms, while also taking advantage of the unique feedback mechanisms afforded by modern computing technology. While sometimes marked by William Higginbotham's 1958 *Tennis for Two*, created at Brookhaven Labs using an analog computer connected to an oscilloscope, most scholars attribute the first digital computer game to Steve Russell, who wrote *SpaceWar!* on a PDP-1 mainframe computer in 1962. The first text adventure game, *ADVENTURE*, was created by William Crowther and later refined by Don Woods in 1975. Both a game and a simulation, *ADVENTURE*was inspired by the Mammoth Cave system in Kentucky, which Crowther and his wife regularly explored. Now, over six million players quest through the vast reaches of Azeroth in the online multiplayer game *World of Warcraft*, while others learn of redemption in the single-player *Prince of Persia* trilogy. With an over forty year history of their own, computer games, their rise, complexity, and popularity, have reminded us that games – the long history of games – are perhaps our most understudied form of art and communication.

The recent prominence of games has led educators and policy makers to turn (once again) to interactive gaming media as a way to educate students and influence opinion, resulting in titles like *Darfur is Dying*, the language and culture military education game *Tactical Iraqi*, the educational game *Discover Babylon*, or the social critique of Kinko's worker in *Disaffected*. Taking games seriously requires understanding their mechanisms, both in order to exploit strategies for successful

education but also to provide suitable, knowledgeable critique of an increasingly prevalent rhetorical platform. Through this continuous process of understanding these models for interactive rhetoric, entertainment, and education, scholars also can begin to build a framework that allows for the preservation for the many digital media forms that are too slowly finding their way into our public archival institutions (with a few notable exceptions).

Games deserve our attention not only because of their potential in education, and not simply for their role as a cultural, historical artifact, but also because they are perhaps the computer-as-medium's most unique art, one that takes the basic function of computing – interactivity – and refines it into an expression, a meaningful exchange often articulated by a challenge between opposing interests. I conclude by providing one example of how to 'take games seriously' as an object of critical study based on its own merits. I provide an overview of 'game fiction,' which I define as a sub-genre of game that draws upon and uses narrative strategies to create, maintain, and lead the user through a fictional environment. The relationship of computer games to more traditional forms, and especially narrative, has been a contentious issue within the growing field of game studies. Understanding games' indebtedness to and departure from other expressive works, particularly the narrative forms of prose fiction and film, proves a delicate challenge. The solution, I argue, stems partly from creating an understanding of genre within games that addresses underlying mechanisms as much as subject matter, and moves beyond the genres that heretofore have been casually outlined. With the example of 'game fiction,' I suggest a framework for studying games – and for distinguishing *types* of games that are all too often lumped together under a singular rubric – that both draws on existing scholarly practice while also taking into account the unique computational framework of the modern computer game.

ALLC Panel: Digital Resources in Humanities Research: Evidence of Value

David Robey *(d.j.b.robey@reading.ac.uk)*
University of Reading
Harold Short *(Harold.Short@kcl.ac.uk)*
King's College London
Thornton Staples *(tls@virginia.edu)*
University of Virginia
Geoffrey Rockwell *(georock@mcmaster.ca)*
McMaster University
Sheila Anderson *(sheila.anderson@ahds.ac.uk)*
UK Arts and Humanities Data Service

While most of us who do humanities computing need no convincing of its value, academic colleagues - including those on appointment and promotion panels - still need to be convinced, and even more so funders. If we want backing for the use and further development of digital resources, both data and processes, we need to collect more extensive concrete evidence of the ways in which they enable us to do research better, or to do research we would not otherwise be able to do, or to generate new knowledge in entirely new ways. Since the value of humanities research as a whole is qualitative, not quantitative, it is qualitative evidence in particular we should be looking for: digital resources providing the means not simply of doing research, but of doing excellent research.

One major task is therefore to accumulate a body of evidence of what researchers have done, and achieved, with digital resources— evidence that is not always as easy to find as one would wish. But what researchers do with these resources is also dependent on the nature and quality of the resources themselves: their academic rigour and completeness, and their technological design. Where a new digital research resource is created, to what extent does an assessment need to focus on the methods of analysis, design and construction of the resource, and to what extent on its usefulness to the community or communities of researchers? How important are the technical standards adopted? What about sustainability? What are the opportunities for re-use and for developing research materials that can be re-configured in a variety of ways?

We therefore need not only to accumulate evidence of value, but to think more about the criteria and methods we should use for doing so. For instance, where an analytical method is based on the methods used in other disciplines, e.g. statistics, there may be methods of assessment that can be likewise borrowed. Many of us believe the evidence is there, but we do not have the range of coherent, agreed and tested methods of assessment that will enable us to make assertions of value that will be reliable and will carry weight.

The panel session will seek to identify the range and kinds of evidence that exists, the methods of assessment that are needed, and the methods of re-use and reconfiguration that can give assurances of value.

Contributors:

David Robey (convenor): Introduction

Harold Short : Examples and case studies

Thornton Staples : Design and management for flexible use

Geoffrey Rockwell : Text analysis

Sheila Anderson : e-Science

Text Analysis Portal for Research, Using the Public Release

Geoffrey Rockwell *(georock@mcmaster.ca)*
McMaster University
Stéfan Sinclair *(sgsinclair@gmail.com)*
McMaster University

Introduction

In April of 2007 the TAPoR project, after extended testing and two public beta versions, released Version 1.0 of the Text Analysis Portal for Research.[1] This portal gathers text analysis tools configured as web services into an environment where users can define the texts they want to work with, select the tools that best work with different text formats, and log their research. The portal is, on the one hand a broker for piping electronic texts to tools and then managing the results, and on the other hand is a study environment where projects can be managed over time. Our poster will do three things:

- Show an overview of how the portal works with annotated screen shots.
- Show the life of the project including the user testing phases.
- We will have a laptop in order to demonstrate a live version of the portal to interested.

Demonstration

Provided we can have a live internet connection we will be prepared to demonstrate:

- How the portal can be used even without an account using the Try It feature or myLinks pages where users can publish sets of texts for analysis.
- How to get a free account from TAPoR.
- How to define texts that may be elsewhere on the internet or uploaded. How to organize these texts using tags and how to publish a set of searchable texts for students or colleagues.
- How to use the Workbench to organize a set of texts and specific tools for a project. How to run tools on texts and how to save results.

- How to log the progress of a research project and share it through TAPoR.

If a live connection is not available we will have a local version running based on the TAPoR Live CD which is a bootable version of linux and the portal that works locally.

User Testing and the Life Cycle of a Tool Project

Version 1.0 of the portal is the result of extensive user testing which will be visually represented in the poster as will ways new users can provide us feedback as we develop the priorities for further releases. [2]

For those familiar with earlier versions of the portal there are a number of improvements including:

- A new type of news channel called a Research Log. Users can save results from tools to this log along with comments. They can also save the text/tool configuration to run again.
- A new feature called Analyze Text that will open a window with the text on the right and suggested tools on the left for close study of one text.
- The myLinks feature has been adapted so users can publish their public texts for others to search across and for other to analyze. This can be used, for example, for providing students with a study set of texts.
- The interface has been simplified to make it easy for users to get projects going.

Background

The TAPor Portal is one of the outcomes of a 5 year project funded by the Canada Foundation for Innovation (CFI) that involves 6 universities across Canada, including the University of Victoria, the University of Alberta, McMaster University, the University of Toronto, the Université de Montréal and the University of New Brunswick. Provincial funding bodies and our respective universities have also supported the TAPoR project. The portal was developed by Open Sky Solutions with a team primarily at McMaster University. The poster will recognize the supporting institutions but will focus on the portal and not present the wealth of other projects that are part of TAPoR. [3]

This poster will show the current version of a major text analysis tools development effort that developed a model for sustainable and modular text analysis for a broad community. This is a challenge that has concerned computing humanists since the 1980s. As Susan Hockey put it in a post to HUMANIST in 1996 that reported on a meeting she convened at CETH, "For some time, those of us active in humanities computing have felt the need for better and/or more widely accessible text analysis software tools for the humanities. There have been informal discussions about this at a number of meetings, but so far no substantial long-term plan has emerged to clarify exactly what those needs are and to identify what could to be done to ensure that humanities scholars have readily-available text analysis tools to serve their computing needs into the next century."[4]

The TAPoR model meets a number of the objectives described for text analysis tools in the 1990s by Hockey and others.[5] It is widely accessible. It provides a long-term model using web services to bring tools together where they can be used in a study environment. It allows new tools to be added and existing tools to be improved or replaced over time as new needs emerge. It allows research to be recorded and shared.

Future Plans

The TAPoR Portal also has limitations. Depending on tools provided as web services by others increases the chances of simple operations not working. Many of the tools are not multi-lingual. The tool broker model where the portal gets a remote text and passes it to a remote tool is not as efficient as an all-in-one model where texts can be preindexed. And, as with any large software project there are still inconsistencies, awkward interface panels, and bugs.

The poster will conclude by outlining future plans. TAPoR hopes to be funded for a second phase through the CFI Leading Edge Fund in order to add large-scale text datamining and aggregation tools. Currently the portal model does not pre-index large collections (except in limited cases) and provides limited aggregation or crawling tools. TAPoR 2, if funded, will continue improving the interface and will add large-scale functionality by adapting crawling, scraping and data mining tools to the study space.

---

1. For more about the TAPoR project see <http://www.tapor.ca>
2. TAPoR interface at <http://www.tapor.ca/interface> details a persona oriented investigation conducted by Audrey Carr and Joanna Dacko. Dr. Wendy Duff at the University of Toronto also conducted extensive interviews that have not been published, but which were absracted for the portal development team. Finally, we have a layered testing process combined with the gathering of tool statistics that is being conducted on the release.
3. *Mind Technologies* edited by Raymond Siemens and David Moorman (U of Calgary Press, 2006) includes a number of articles about the variety of projects supported by TAPoR.

4. Susan Hockey, *Humanist Discussion Group* 10.54 (23 May 1996). See <http://lists.village.virginia.edu/lists_archive/Humanist/v10/0054.html>. There is also a note from Michael Sperberg-McQueen pointing to a trip report available at <http://tigger.uic.edu/~cmsmcq/trips/ceth9505.html>
5. Geoffrey Rockwell and John Bradley, "Eye-ContTact: Towards a New Design for Text-Analysis Tools", CHWP A.4. (1998). See <http://www.chass.utoronto.ca/epc/chwp/rockwell/>

Recent Developments in the Music Encoding Initiative Project: Enhancing Digital Musicology and Scholarship

Perry Roland *(pdr4h@virginia.edu)*
University of Virginia
J. Stephen Downie *(jdownie@uiuc.edu)*
University of Illinois at Urbana-Champaign

Introduction

The MEI (Music Encoding Initiative) is an encoding toolset meant for modeling music information (Roland, 2006). This does not refer to audio, but to the symbolic representation of Common Music notation (CMN), the “standard” form of music notation used between 1700 and 1935 (see Fig. 1). MEI is an XML application expressed in DTD form (for the present), with a TEI-like header (Fig. 2), a `<music>` element instead of `<body>`, and `<front>` and `<back>` to accommodate text. The goals for MEI are to encode CMN “out of the box”, to limit verbosity without compromising the self-documenting aspect of XML, to support repertoires other than CMN, and to support creation of multi-lingual interfaces by allowing generic identifier names to be changed, all to better enable creation of scholarly editions. In this poster presentation we highlight the recent developments made to improve and extend MEI with a special emphasis on the impacts these developments can have on digital musicology and scholarship.

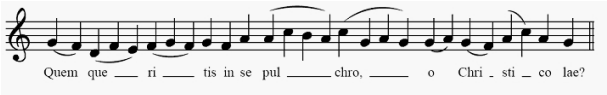


Figure 1. Sample CMN rendition of "Quem queritis"

```
<?xml version="1.0" encoding="UTF-8"
standalone="no">
<!DOCTYPE mei SYSTEM
"http://www.lib.virginia.edu/digital/resndev/mei/mei17b/
mei17b.dtd">
<mei version="1.7b">
<meihead>
<meiid/>
<filedesc>
<titlestmt>
```

```
<title>"Quem queritis"</title>
</titlestmt>
<notesstmt>
<bibnote type="encoding-date">2003-03-15</
bibnote>
</notesstmt>
</filedesc>
<profiledesc>
<langusage>
<language id="la"/>
</langusage>
</profiledesc>
<revisiondesc>
<change>
<changedesc>
<p>Transcoded from MusicXML version 1.0</p>
</changedesc>
<date>2006-09-26-04:00</date>
</change>
</revisiondesc>
</meihead>
```

Figure 2.Sample of the TEI-like MEI header for "Quem queritis".

Recent Developments

MEI is extending its notation encoding capacity beyond CMN. While most representations of notation have been limited to CMN, MEI has a TEI P4-like extension/restriction mechanism which can be used to expand the universe of sources to which it can be applied. Work is progressing in two particular areas: “White mensural” notation and “Medieval neumatic” notation. Both of these developments also better situate MEI as an encoding mechanism for serious digital musicology (see Section 3).

Mensural notation is a musical notation system used from the later part of the 13th century until about 1600. The name “mensural” refers to the capacity of this system to notate complex rhythms with great exactness and flexibility. Mensural notation was the first system used for European music that systematically used individual note shapes to denote temporal durations. Measure-less music requires structural reorganization, or score-by-staff encoding rather than score-by-measure-by-staff encoding: the staff must directly allow features previously available only within `<measure>`, such as for timed events. MEI is introducing new concepts to enable such encoding: events -- ligature, mensuration and proportion signs; orthographic vs. semantic accidentals; barlines, and new attributes and attribute values.

Neumes are the basic elements of Western and Eastern systems of musical notation prior to the invention of five-line staff notation. Neumatic notation eventually evolved into modern musical notation, but remains standard in modern editions of plainchant. MEI is introducing new concepts to enable encoding of this system: single and compound neumes, ligatures, mensuration and proportion indicators, orthographic accidentals, custos, interpretative marks (episema, liquescent neumes, quilisma) that can be treated as timed events, mora (an augmentation dot) that may be handled the same as a modern augmentation dot, and new attributes and attribute values.

MEI is now supporting data collection via MusicXML (Good, 2002), a translator and interchange format for common Western musical notation from the 17th century onwards. A new 2mei XSL style sheet facilitates transformation of MusicXML to MEI, creating an input path from any software that supports export of MusicXML. On average the resulting MEI files are only 51% as large as the original MusicXML files.

Enhancing Support for Digital Musicology and Scholarly Editions

MEI has been gaining favorable reception among those engaged in digital music scholarship. For example, the MeTAMuSe project has said, "In contrast to MusicXML, which is the de facto industry standard, but which is rather limited in the representation of musicological concepts such as multiple divergent sources, MEI has definite advantages in the musicological context" (Byrd et al., 2006). MEI has also been described as one of "two really serious contenders" in this problem space (Kay, 2004).

Digital critical editions of music start with the encoding of the musical sources, and then add layers for presentation and meaning. MEI's improved capacity to encode multiple types of notation, provide support for non-transcriptional text commentary/annotation, as well as improved methods of data capture into the encoding format and data export will hopefully foster new efforts to create scholarly critical editions using XML. Such content-based encoding modeled on text encoding formats is not only best-suited to the development of these digital editions, but can potentially best document the intellectual process of the development of the corpus, making the critical work better suited for verification and scholarly argument.

MEI is well-suited for the creation of scholarly editions that document the creation and revision history of a single musical composition. In MEI, a single file supports encoding of the data common to all sources only once, rather than requiring redundant markup and encoding in multiple files. The `<source>` element holds bibliographic and physical

description of a single source document and can be linked to specific data via its data attribute, while data can be linked to the source via the source attribute, a mechanism not unlike the "declarable/ declaring" attributes in TEI. When the meiCrit parameter entity is enabled, parallel alternative encodings are possible at the score, measure, and staff levels. This feature would be particularly useful for the construction of such scholarly editions as the "Online Chopin Variorum Edition" (<http://www.ocve.org.uk/>). In the case of manuscript music, the <handlist> element in the header and the "hand" attribute (available on most music content elements) allow one to track the scribes, copyists, etc. who notated the music.

MEI now supports non-transcriptional text commentary/annotation. The <annot> element provides a way to group participating events, the notes that form a descending bass line, for example, and provide a label for the group. An editorial or analytical observation, encoded elsewhere, may be pointed to using the linking attributes. Alternatively, the observation may be included directly within the <annot> element.

Ongoing and Future Work

Work is progressing to support two modes of visualization and export. The first is a conversion from MEI to MusicXML which will allow any software that reads MusicXML to display/manipulate the data. This is analogous to the method, i.e., conversion to HTML which is used to display SGML/XML. This MEI-to-MusicXML conversion, is, however, lossy. The second mode, direct conversions to other internal representations, requires writing a filter for each existing data format. This is a time-consuming task, but will greatly reduce data loss.

Also, additional elements necessary for manuscript encoding are planned: <add> (for something added by another person or at a later date), <del> (for something marked out), <unclear> (for illegible passages), <damage>(damage to the carrier), <supplied> (for data supplied by the editor), and <handshift> to indicate a change in scribal hands.

Appendix

```
<work>
<music>
<mdiv>
<score>
<scoredef>
<staffgrp>
<staffdef n="1" id="P1" label.full="Voice"
clef.line="2" clef.shape="G" midi.div="2"/>
</staffgrp>
</scoredef>
<section>
<measure n="1" id="d1e18" right="dbl">
<staff def="1">
<layer def="1">
<note  id="d1e32"  tstamp="0"  pname="g"
oct="4" dur="4" dur.ges="2" stem.dir="up">
<verse n="1">
<syl>Quem</syl>
</verse>
</note>
<note  id="d1e53"  tstamp="2"  pname="f"
oct="4" dur="4" dur.ges="2" stem.dir="up"/>
<note  id="d1e69"  tstamp="4"  pname="d"
oct="4" dur="4" dur.ges="2" stem.dir="up">
<verse n="1">
<syl wordpos="i" con="d">que</syl>
</verse>
</note>
<note  id="d1e90"  tstamp="6"  pname="f"
oct="4" dur="4" dur.ges="2" stem.dir="up"/>
<note  id="d1e104"  tstamp="8"  pname="e"
oct="4" dur="4" dur.ges="2" stem.dir="up"/>
<note  id="d1e120"  tstamp="10"  pname="f"
oct="4" dur="4" dur.ges="2" stem.dir="up">
<verse n="1">
<syl wordpos="m" con="d">ri</syl>
</verse>
</note>
{CODE DELETED FOR SPACE}
</measure>
</section>
</score>
</mdiv>
</music>
</work>
</mei>
```

Bibliography

Byrd, Donald, Time Crawford, and Geraint Wiggins. MeTAMuSE: Methodologies and Technologies for Advances Musical Score Encoding. Project proposal submitted to The Andrew W. Mellon Foundation. Final Version, 19 May 2006, Typescript. 2006.

Good, Michael. "MusicXML in Practice: Issues in Translation and Analysis." *Proceedings of First International Conference*

MAX 2002: Musical Application Using XML, Milan, September 19-20, 2002. 2002. 47-54.

Kay, Michael. *XSLT 2.0 Programmer's Reference*. 3rd edition. Indianapolis, IN: Wiley, 2004.

Roland, Perry. *The Music Encoding Initiative (MEI)* . 2006. <http://www.lib.virginia.edu/digital/resndev/mei/>

Multilevel Displays and Document Blueprints: Dynamic Browsing Using XML Structures and Text Features

Stan Ruecker *(sruecker@ualberta.ca)*
University of Alberta
Stéfan Sinclair *(sgsinclair@gmail.com)*
McMaster University

Abstract

In this paper, we discuss how researchers can benefit from tools that allow them to work with visualizations that rely on XML data at different levels of granularity. For these visualizations, we propose using a single interface that draws on the underlying structural information at both the collection level and the level of the contents of the individual documents. We compare two models of this kind of interactivity that are the subjects of our current interface design and prototyping activities. One of these systems is predicated on displays relying on sequence, and the other on visualizing the structure of items and facilitating their traversal.

Discussion

The common goal of these interface systems is to provide researchers with experimental means of combining overviews of a document with tools for manipulating the display. First is our multilevel document visualization system (Ruecker et al. 2005) that combines three or more simultaneous displays, including a microtext column showing collection items, a separate microtext column showing the contents of a particular document, and a reading view of the document. These different levels of display are combined with tools for selecting and manipulating a portion of the text in a subsequent display.

The purpose of the multilevel system is to allow the user to work with a digital document within the visual context of related material. We have currently installed it as a feature in the prototype for watching digital scripts, where the standard reading view is supplemented with a digital stage that allows the reader to see both the dynamic text playback and the blocking of the characters as they move around the stage. We

are currently carrying out a user study of actors using the system to learn their lines.

In addition to the dynamic text, we have recently begun to consider the possibilities of using the microtext columns as opportunities for providing overviews that can provide further data. For example, they might be colour-coded and re-organized according to some useful principle. In the Watching the Script demo, one form of colour-coding currently available – for plays that have been encoded in XML that marks the character names – is to help the reader differentiate between characters.

For purposes of a director planning the play, this overview allows exploration of the various relationships of the characters on the stage. It can be use to address questions such as who is on stage when, and with who, and how many lines does each actor have? By selectively applying colour only to the character or characters of interest, and having the remainder display in regular black text, the director has a tool for studying the entire play from the perspective of the staging.

Another overview panel in the prototype provides a list of the characters. As a simple display it does not provide many affordances, but with the addition of some further information and the ability to sort, this list also has potential benefits. For example, if we give the user the opportunity of adding next to each character the number of lines that character speaks, or a total of time spent on stage, or both, and then rearranging the characters according to those numbers, we have another tool for use in understanding the play at an overview level, and for planning production. Combined with the blocking tool, these overviews and their related tools begin to create a tool suite that we hope will prove useful to actors, directors, and students of theatre.

The second prototype produces what we are calling a Document Blueprint: a compact visual representation of the markup in a document. This blueprint can be used to suggest encoding particularities of a document by colour-coding tags and attributes in flexible ways. The identity – and therefore colour-coding – of sections can be defined in ways that ignore certain nodes in an XML tree, including elements, attributes, and text.

Our first application of this system is to generate a document view where the table of contents is initially displayed in a legible font size and the markup of sections are displayed in small, blueprint mode (where visual cues from the text emerge but the text itself is too small to be legible). The Document Blueprint system allows the user to define which elements to toggle between legible and blueprinting modes. Depending on which elements the user chooses, the system provides a variety of overviews that can help both in understanding and navigating the document, much as the conventional table of contents has always done, but with the important differences of dynamic display and interactivity (Ruecker 2005).

Future Directions

For the multilevel display, we are currently working to extend the affordances of the various overview items. It may be useful, for instance, to allow the entire surface to be temporarily subsumed by the overview that currently only occupies one column. The more complex display would show the data, still not at a reading view, but at a smaller level of granularity. Colour-coding and reorganizing this form of display could prove both interesting and useful.

The document blueprinting project is the youngest of our prototype projects and will likely evolve the most in the coming year. We are developing this system in collaboration with the Orlando Project, which in its next phase will be producing literary critical material in volume form. The dynamic table of contents in this case will be used as a navigation aide for readers interested in browsing through the extended prose of the three volumes. As has been the case with each of the prototypes, we will pursue an iterative cycle of design, prototyping, and development, with user study and experiment at each stage of the cycle.

Bibliography

Ruecker, Stan. "The Electronic Book Table of Contents as a Research Tool." President’s Panel. Congress of the Humanities and Social Sciences: Consortium for Computers in the Humanities / Consortium pour Ordinateurs en Sciences Humaines (COCH/COSH) Annual Conference. London, Ontario. May 27- 30, 2005. 2005.

Ruecker, Stan, Eric Homich, and Sinclair, Stéfan. "Multi-level Document Visualization." *Visible Language* 39.1 (2005).

Sinclair, Stéfan. "Computer-Assisted Reading: Reconceiving Text Analysis." *Literary & Linguistic Computing* 18.2 (2003): 175-184. doi:10.1093/llc/18.2.175

Twelve Hamlets: A Stylometric Analysis of Major Characters' Idiolects in Three English Versions and Nine Translations

Jan Rybicki *(jrybicki@ap.krakow.pl)*
Pedagogical University

Following my own comparative stylometric analyses of originals and translations of literary works (Rybicki 2006, 2006a), I have decided to expand the scope of this research to more than just two languages. This has been made possible by the fact that Burrows's well-established method, first used in his study of Jane Austen (1987), and later developed, evaluated and applied by a number of scholars, including Hoover (2002), can be applied to the most frequent words of any text, also in a language unknown (or less-known) to the researcher. Quite self-evidently, Shakespeare's *Hamlet* was chosen for its status of a crucial work of English literature and its numerous translations. The fact that the English *Hamlet* exists in three primary versions (First Quarto, Second Quarto, First Folio) was also a serious incentive.

Character parts were extracted from the three above-mentioned originals and nine translations: Czech (by Josef Jirí Kolár, 1855), French (François-Victor Hugo, 1863), German (August Wilhelm Schlegel, 1798), Hungarian (János Arany, 1864), Italian (Goffredo Raponi, 1999), Polish (ca. 1875), Portuguese (1966) Russian (Mikhail Morozov, 1954), and Spanish (Leandro Fernandez de Moratin, 1798). The selection of the translations was a compromise between their age (the older the better) and availability in electronic form. The three originals were taken from the collection of exemplary electronic texts on TACT CD (1996). Most frequent words were identified in each version, basing on the rule that a given word was allowed if it appeared in at least 5 of the major character parts: these included Claudius, Gertrude, Hamlet, Horatio, Laertes, Ophelia and Polonius. As a result, the number of the most frequent words included in the analysis varied from 157 (First Quarto, obviously the shortest text) to 282 (the Italian translation). Relative frequencies were then obtained for each character's use of each of the above-mentioned words; matrices of the relative frequencies were then used to produce multidimensional scaling (MDS) graphs for each version of the play. The same material was also treated with cluster analysis, usually with good agreement with MDS.

The following observations have been made:

1. Graphs for each version of the play always included at least two peripheral data points, invariably including the two female characters, Gertrude and Ophelia;
2. Laertes and Horatio frequently joined Gertrude and Ophelia in their peripheral orbits around the data point for Hamlet, usually central in all graphs (with the notable exception of the Czech translation);
3. In many translations, the idiolect of Hamlet was more or less similar to those of some other characters: its closeness to Polonius (or Corambis) in the First Quarto lessened in the Second, only to be replaced by a much stronger similarity to Laertes. An even stronger similitude was observed between Hamlet, Polonius and Claudius in the Portuguese translation, or to Claudius alone in the Italian version. In general, Polonius and Claudius were the idiolects most consistently similar to that of the main character.
4. The closest similarity of pattern between entire graphs was that between the First Quarto and Schlegel's German translation (Hamlet and Polonius surrounded by even-spaced peripheral characters);
5. The general pattern was also quite similar between the Second Quarto and the Folio (peripheral Gertrudes, Horatios and Ophelias, central Hamlets, other characters in-between); it was also roughly reproduced in the translations by Arany, Moratin and Paszkowski.
6. The clearest gender division was observed in the Czech, Hungarian, and Russian translations.

This introductory study seems to offer several promising avenues for development. The similarity of pattern between the Second Quarto and the Folio is not surprising; even less so is its above-mentioned reproduction by three translations, since their source material was usually a compilation of the two original versions. On the other hand, Schlegel's similarity to the First Quarto seems pure coincidence, as the so-called "bad quarto" was discovered more than twenty years after the early German translation. The peripheral position of female characters in all graphs is perhaps the most consistent feature of most of European writing so far analysed with MDS; this effect also travels very well in translation.

Yet the greatest potential for expanding this study lies in the sheer number of *Hamlet* translations. In the languages included in this study, they are at least five, and very often more than ten; cultures such as Polish, German, or French, have produced almost twenty *Hamlet*s each. The inclusion of at least some of these would provide a fuller picture of patterns of stylistic differences between Shakespeare's fundamental play and its various realisations in other languages.

Bibliography

Burrows, J. F. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press, 1987.

Hoover, David. "New Directions in Statistical Stylistics and Authorship Attribution." *ALLC/ACH 2002 Conference Abstracts.* Tübingen: Tübingen University, 2002.

Rybicki, Jan. "Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations." *Literary & Linguistic Computing* 21.1 (2006): 91-103.

Rybicki, Jan. "Can I Write like John le Carré?" *Digital Humanities 2006 Conference Abstracts.* Paris: CATI, Université Paris-Sorbonne, 2006a.

GRADE: a GRAmmar Development Engine

Harry Schmidt *(hgschmid@princeton.edu)*
Princeton University
Helma Dik *(helmadik@uchicago.edu)*
University of Chicago

The Grade project allows authors of complex reference works to encode documents using familiar features from print reference works, but, crucially, adds the possibility for the author to mark where For_'Dummies'-level information stops and For_The_Dozen_Living_Experts-level information starts, plus the relevant stages in between. In turn, readers of Grade documents are able to switch from 'Dummies' mode to Expert mode at the click of a button. The back-end to all of this is provided by an XML database. Open-source release of Grade 1.0 is planned to coincide with Digital Humanities 2007.

Rationale

Existing grammars of any language are not easy to use and typically envisage only one kind of user at a time. But a user may vary in her preferences from moment to moment. Does she want 'just the rule' right now? Does she want to see examples, cited in the original language and in translation, and perhaps print them out? But tomorrow, perhaps, she also wants the evidence for that rule, or discussion of its problematic status as a rule?

Any printed reference grammar, even one for a homogeneous audience, is inevitably a product of compromises. Space devoted to discussion of different schools of thought detracts from clarity; space devoted to exceptions detracts from main rules; space devoted to examples and translations clutters the page, etcetera. Paradoxically, then, every good grammar is a bad grammar precisely for the reasons that it is good. A solid number of sensibly discussed examples coupled with a fair representation of scholarly opinion on an issue will never fail to obscure what should be the statement of a simple rule; a simple statement of a rule, however, will never do justice to the complexity of the language and needs, at a minimum, illustrations from actual texts.

An electronic reference grammar, on the other hand, does not have to make any of these painful choices. What it can offer instead is the capability for any end user to select, with the click

of a button, for any part of the syntax, the level of detail he or she wishes to see at that particular moment, and to switch at any time to further detail, extra examples, or, conversely, back to only the main rule, with no exceptions, no examples, and no translations of examples. If desired, users can choose for themselves what section to print, and in what format (through application of XSLT).

Current collections of thematically unified materials, such as The Perseus project (perseus.tufts.edu) offer user configuration in the form of cookies, which set a user's default language (original texts or translations), default font, and the like. However, Perseus and collections like it derive most of their power not primarily from this user configurability but from the intelligent linking of resources that were previously separate: Words in the texts can be analysed by the morphological parser, which in turn provides links to the lexicon, which in turn allows for further word searches, etcetera.

The Perseus texts, then, can be said to 'talk to each other' and enhance the usability of resources compared to their print versions: The system is aware, for instance, that a user is reading Thucydides when she clicks on a word to get a parse or a meaning; accordingly, the system will highlight Thucydides citations in any dictionary entry called up. GRADE takes this configurability one step further in allowing users to retrieve information customized not just for the text that they are reading but for the level of complexity that they select.

While traditional rich encoding following, e.g., TEI guidelines has typically aimed for faithful transcription of an original print publication, or generally, has aimed to serve the intents of an author/publisher, the current project uses its XML schema to allow the end user this flexibility at all times, so that this one grammar can become, in effect, all grammars to all people. At the same time, GRADE empowers the author in all the ways that electronic publishing has started to empower authors, and then some: a GUI for authoring, the possibility for piecemeal and immediate publication, as found in blogs and wikis, but paired with a capacity for richly structured data and full editorial oversight.

GRADE finds its origin in the second author's need for a grammar authoring tool, but we submit that the system can be re-used for many projects that share the characteristics of complexity and an intended audience with widely varying goals and levels of expertise. Accordingly, in this demonstration of the beta version, a second humanities application is offered in the form of a Latin learners' vocabulary which is organized in sets of basic vocabulary and author-specific vocabulary, and in which database translations from Latin into several languages can be stored. Users can opt for display of the vocabulary in Latin and any of these languages, can select particular authors or genres, can opt for display in flashcard format, etc.

Design Features

GRADE is written in Perl, chosen for its robustness and because it is a well-supported Web-centric application programming framework. GRADE is written in such a way that even a novice computer programmer can write extensions that increase its functionality. Two templates (grammars and vocabularies) will be available in June 2007; tutorials on extensions are foreseen for Summer 2007.

GUI for XML database

The project has a graphical user interface for its XML database, which allows for entry, update and management of the objects (grammar sections) and their hierarchical organization (grammar chapters and smaller sections) in the database and for their encoding as specific types of information: text can be marked as examples, translations, notes, levels of complexity, tables, and lists. Authors can enter text in a standard browser (Firefox) without having to add in XML tags by hand (figure 1:editor screen). This makes for convenience for the authors but also avoids validation problems. The native XML database backend is based on Sleepycat Software's open-source Berkeley DB XML 2, provides full support for Unicode UTF-8 and allows queries via XPath 2.0 and XQuery 1.0. Standard open-source tools are used throughout, including Apache and Perl.

Rendering Engine

The rendering engine gets objects from the database (sections of the document) and renders them with XSLT. Users (that is, readers of the grammar) can browse through different levels of complexity, choosing to display more or less information, and what types of information to display, with a click on 'collapse boxes' or by adjusting their preferences (figure 2: Rendering options).

GRADE supports common Web formats and protocols for information exchange (XML, XMLRPC, and JSON), so that another project could potentially even integrate a real-time data feed (that is to say, up-to-date data rather than a particular revision which may become obsolete). Alternately, "calling" programs can also request a specific revision of a document to ensure its stability.

User registration and privileges

Users on the authoring side need administrative privileges to edit the database (see further below). For non-author

users, cookies will retain preferences (for display level, translation display, etc.) during one session. Users may choose to register in order for preference to be retained beyond sessions. The user database recognizes different levels of privileges for users. It permits those with administrative privileges to make changes to the database at will, subject to hierarchical controls that prevent changes to sections controlled by higher-order administrators. A versioning system, at the bottom of the editing screen, guarantees the integrity of all documents in the system and permits users to compare prior edits of a given section. In its final state, GRADE will provide the kind of precise revision tracking necessary for a rigorous academic project, including one feature that all existing Wiki software (including MediaWiki, the software used by Wikipedia) lacks -- the ability to see, from within a document, who wrote what portions of it and when.

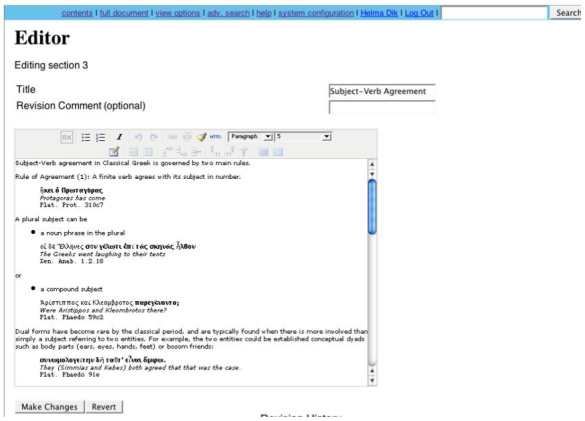


Figure 1: Editor screen. Note 'depth level' 5 next to 'paragraph'.

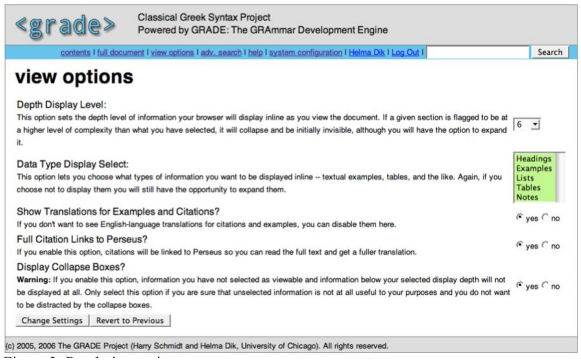


Figure 2: Rendering options

What GRADE is not

GRADE is emphatically not a pedagogy application or learning management system, a la SCORM from ADL. It empowers both authors, by allowing them to present richly coded complex content, and end users, by giving them a flexible document to use and explore, not by feeding them content piecemeal or by tracking them. An internet-savvy audience, we submit, now long-accustomed to browsing hypertexts, should not be forced into SCORM's 'sequencing,' which aims to have users experience content in previously specified order.

Conclusion

GRADE enhances electronic publishing formats in important ways. It offers an important toolkit for authors in the Humanities and beyond that will allow them to empower the readers of their texts and thereby reach wider audiences. GRADE allows for co-authoring without jettisoning editorial oversight or losing track of who wrote what; it provides a powerful tool both for authors in fields in which a rich tradition of grammatical description already exists, such as many European languages, as well as for those in need of a tool for writing the first reference grammar of a language ever to be published.

The Versioning Machine 3.0: Lessons in Open Source Software [Re]Development

Susan Schreibman *(sschreib@umd.edu)*
University of Maryland Libraries
Ann Hanlon *(ahanlon@umd.edu)*
University of Maryland Libraries
Sean Daugherty *(seancdaug@umd.edu)*
University of Maryland Libraries
Tony Ross *(tonyross@umd.edu)*
University of Maryland Libraries

The Versioning Machine made its debut at the 2002 ALLC/ACH Conference in Tübingen. It is a software tool for displaying and comparing multiple versions of texts designed by a team of programmers, designers, and literary scholars. A primary goal of the software was to create a display environment which provides for features traditionally found in codex-based critical editions – such as annotation and introductory material – that also takes advantage of opportunities offered by editing and displaying witnesses in an electronic environment.

The VM was designed as open source software that would allow literary editors to concentrate on editing and displaying multiple versions of text using the Text Encoding Initiative (TEI), rather than expecting them to build a custom environment. The developers deliberately built the tool with JavaScript, CSS, and XSLT, as these technologies are presumably within the reach of a humanities scholar if she wanted to alter the environment.

The Versioning Machine is now in its third iteration, and has been demoed at various humanities computing conferences over the years. This poster session will not focus on the VM as a tool per se, but will focus on the issues and lessons learned developing open source software for the humanities. As this tool was developed for a very specific goal for a potentially small community of users, this poster will focus on the issues raised by developing software for an emerging practice, the difficulties posed by changing technologies, and the issues raised by moving from fairly standard book-based presentation paradigms for scholarly editions, to the relative vacuum of agreed-upon conventions for web-based editions

Developing Software for an Emerging Practice

At any gathering of digital humanists, there is a plea for software designed specifically for humanities applications. *The Versioning Machine* was developed to fill a fairly narrow niche. As such, its development has had as a central focus the exploration of the possibilities, as well as the limitations, of electronic editions that focus on the presentation of multiple witnesses.

For the first time in the software's development, rigorous and methodical usability testing will be carried out by a team member not directly concerned with software development. This testing will not only assess user-friendliness, accessibility, performance, and overall structure, but will also investigate the larger issue of whether *The Versioning Machine* functions as an asset to the community of scholars it is meant to serve. The test participants will be comprised of 10-15 literary scholars and textual editors. They will be given a series of tasks to perform and will be asked to comment on what they are doing or trying to do. They will also be asked to comment on the overall difficulty or ease of use, as well as the degree to which *The Versioning Machine* has or could enhance their scholarship. Because many of the participants will not be on-site, a web-based survey will be used to capture their impressions. Additionally, a number of participants will be tested on site so that our observations can be compared with the self-reported experiences. The resulting findings will be used as a framework to make further changes to the software.

Difficulties Posed by Changing Technologies

Persistent issues frequently faced in web-based open source tools development include shifting platforms, technologies, and standards on which software is constructed; personnel changes within the development team; and finding enough time to not only do software development, but create the documentation and procedures which allow others access. The difficulties of attempting to adhere to standards, while ensuring cross-browser compatibility, is one issue that will be addressed in the poster as representative of these issues.

During the early stages of the tool's development, it was not possible to replicate the functionality available for browsers on the PC on the Mac OS operating systems. It was decided to go live with limited browser support rather delay the software further. By 2.0, many of the compatibility issues had been worked out, with support for Mozilla-based browsers and Apple's Safari. 3.0 focused both on expanding features of *The*

Versioning Machine as well as addressing several major compatibility issues. This involved a significant reworking of both the XSLT stylesheet and the CSS style rules that power the tool. Earlier versions were lacking in compliance with W3C standards, which accounted for a large part of the cross-browser compatibility issues. A change in personnel meant that several of the people involved in versions 1.0 and 2.0 were no longer on the project team. A lack of both comprehensive documentation and time ensured that, even in its current form, *The Versioning Machine* would not correctly validate as standards-compliant. Thus a pragmatic approach was adopted in which the most obvious discrepancies were addressed, while ensuring that all new additions were coded with respect to the relevant W3C standards.

Of the newly added features, the most significant was the introduction of optional line numbering, drawn from the TEI markup of the document. This proved one of the more difficult features to implement, in large part because of the different ways that various browsers implement CSS. A presentation that looked acceptable in Internet Explorer, for example, did not look acceptable in Firefox or Safari. Toggling line numbers on and off presented another problem. In this case, the JavaScript required worked differently on all three browsers. This was one example of a problem that strict adherence to W3C standards could not solve. There was more than one standards-compliant way of achieving similar functionality, but the difficulty was in finding one method that worked consistently across multiple platforms.

The ability to directly modify the version 2.0 source code certainly expedited the design process for version 3.0. On the other hand, it also meant that we were forced to address many of the shortcomings of the earlier version as well. Thankfully, Amit Kumar, who had worked with the earlier version and implemented the majority of the version 3.0 changes, was able to provide insight into the workings of the application, and to bring his experience working with it to bear on the new compatibility challenges.

As part of the 3.0 redevelopment, a thorough re-evaluation of the interface was undertaken which would move it toward a more finished, professional look. A common pitfall of interface design is that the person designing the interface is often so familiar with the product that they are unable to view it from a new user's perspective. Bringing in a new designer for release 3.0 meant not only a fresh approach to the visual side, but also a new set of eyes completely unfamiliar with *The Versioning Machine*.

Other than the addition of a "backup" menu to the footer, the interface elements of 3.0 are not significantly different from release 2.0 — however, their arrangement has been streamlined for better usability. Even those changes that appear purely cosmetic have a level of thought behind them. For example, the text in the upper left that said "Versioning Machine" now spells out the release version of the software. To the left of that a tagline ("A Tool for Displaying & Comparing Different Versions of Literary Texts") has been added to spell out in broad strokes the software's intended purpose.

What Does an Electronic Scholarly Edition Look Like?

Lastly, this poster will address some of the theoretical issues that the developers of *The Versioning Machine* have faced in designing an environment to present a web-based scholarly edition. There is today, ten years after the development of the World Wide Web, little consensus within the editing community about the features and standards that should be required, or at least desired, in these editions. The user testing on *The Versioning Machine* to be presented at *Digital Humanities 2007*, the ongoing development of the software to keep apace with standards, and changing user expectations, will contribute a small piece to that dialogue.

Bibliography

Better Desktop. Accessed 2006-11-12. <http://www.betterdesktop.org>

Burstein, Cari D. "Viewable with Any Browser Campaign." 2006. Accessed 2006-11-12. <http://www.anybrowser.org/campaign/index.html>

Cockburn, Craig . "Cross Browser Compatibility and Website Design." 2005. <http://www.siliconglen.com/usability/browsers.html>

Kaufman, Joshua. "Practical Usability Testing." 2006. Accessed 2006-11-12. <http://www.digital-web.com/articles/practical_usability_testing/>

Levi, Michael D., and Frederick G. Conrad. "Usability Testing of World Wide Web Sites." U.S. Department of Labor: Office of Survey Methods Research, 2002. Accessed 2006-11-08. <http://stats.bls.gov/ore/htm_papers/st960150.htm>

Nielsen, Jakob. "'Alertbox: Why You Only Need to Test with 5 Users." 2000. Accessed 2006-11-08. <http://www.useit.com/alertbox/20000319.html>

Olson, George. "The State of the Web: Browser Incompabilities Undermine Web's Foundations." 2000. Accessed 2006-11-12. <http://www.webstandards.org/press/releases/2000-state-web>

Petersen, Jeremy. "A Barebones Guide to Usability Testing." . Accessed 2006-11-08. <http://javaboutique.internet.com/articles/Usability/index-3.html>

Schreibman, Susan. "Computer-mediated Texts and Textuality: Theory and Practice." *Computers and the Humanities* 36.2 (2002): 283-293.

Schreibman, Susan. "The Text Ported." *Literary & Linguistic Computing* 17.1 (2002): 77-87.

Smith, David. "Textual Variation and Version Control in the TEI." *Computers and the Humanities* 33.1-2 (1999): 103-112.

Sperberg-McQueen, C. M., and Lou Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange, XML-Compatible Edition.* Oxford, UK: TEI Consortium, 2002.

Meaning and Mining: the Impact of Implicit Assumptions in Data Mining for the Humanities

D. Sculley *(dsculley@cs.tufts.edu)*
Department of Computer Science
Tufts University
Brad Pasanek *(bpasanek@annenberg.edu)*
Annenberg Center for Communication
University of Southern California

In working across and between disciplines, it is the tacit assumptions that may be most destructive to meaningful collaboration. Ours is a state of mutual ignorance, and the goals and practice of the professional literary historian and the machine-learning researcher are equally obscure. But in collaboration mutual ignorance becomes an opportunity for self-reflection, clarification, and the speaking of what is usually unspoken. Willard McCarty writes, “Computational form, which accepts only that which can be told explicitly and precisely” proves “useful for isolating ... tacit and inchoate” knowledge (256). Collaborators are forced to set out a program in detail, one that is mutually comprehensible but also one that delivers results that are simultaneously meaningful in two disciplines. In this paper, we discuss the tacit assumptions that accompany data set preparation, hypothesis testing, and data exploration in order to deliver prescriptive claims. We propose a communication protocol designed to bring hidden and tacit assumptions into plain view where they may be discussed and analyzed. This paper is the third in a series of collaborative efforts undertaken by the two authors. It is informed by real experience working together: working often at cross purposes, garbling a common language, but ultimately producinresults that are of interest to both computer scientists and literary scholars.

Transforming literary content into data for machine learning methods requires the adoption of a number of initial assumptions, each of which significantly impacts the final results. First, the collaborators must select or design an appropriate data representation. The selection of a bag of words model may be one such decision, but other feature mappings such as parse trees or link structure graphs may be more informative for a given task. Once the textual material is represented, we must decide upon a method of feature weighting; that is, we must decide if some features are more

important than others and how much so. Because many learning methods prove intractable when working with very large numbers of features, feature selection is necessary in order to enable computation. Sophisticated, ambiguous, unstable texts must be normalized to make comparisons across texts meaningful—so that the choice of normalization method is critical. So, too, are methods of noise filtering. There are no clear objective choices among methods, because each choice introduces a set of assumptions and biases. We demonstrate these difficulties with experiments on a range of literary data. A loose analogy is here drawn as the literary scholar may choose to cite post-colonial theory to the exclusion of queer theory or practice close reading to the exclusion of historical analysis. We do not argue that structuring assumptions be minimized or eliminated--this is impossible--but we do make the case that in interdisciplinary work especially, it is important for the impact of each assumption to be assessed and reported at the outset. The critic is often a bricoleur, borrowing from literary theory in promiscuous fashion. In preparing data, bricolage is not a ready option and the collaborators must make painful compromises.

The use of machine learning for the testing of literary claims also has several potential pitfalls. The broadest is the impact of the No Free Lunch Theorem, which states that there can be no single machine-learning algorithm that gives optimal performance on all data sets. The choice of a learning algorithm entails, again, the adoption of tacit assumptions about the underlying structure of the data. We may assume that data is linearly separable (which is often a true assumption in text classification), or that the data examples are statistically independent of one another (which is often false in the text domain). As we demonstrate experimentally, these assumptions carry significant impact on the results of the data mining. In the literary domain, selection bias seems particularly problematic as we navigate the politics of canon formation, the difficulty of defining of genre, and the vagaries of influence--all of which trouble the initial selection of texts.

An important question in both machine learning and literature is that of generalization. Do the results and models we discover apply only to our particular data set (as in the case of rote learning), or do these patterns also describe new periods and genres, data we have not yet investigated? In truth, machine-learning methods can never guarantee generalization. However, they do offer statistical bounds on the probability that a model will generalize. According to the Probably Approximately Correct paradigm of computational learning theory, a model that achieves a given level of accuracy on a training data set will likely achieve a predicted level of accuracy on a test data set from the same distribution. The computer scientist emphasizes that generalization bounds are only valid under assumptions of statistical independence in the training data. Care must be taken in the literary domain to ensure that probabilistic assumptions are satisfied. Otherwise, the findings may reflect little more than the selection bias of the investigator. We provide concrete examples of these issues using data from literary analysis, and give guidelines for determining when a generalization assumption may or may not be valid.

The literary scholar often turns to computational methods to explore large numbers of texts--more texts than one human could ever read closely. In this last case, the scholar may not have a hypothesis to test, but is instead looking for new perspectives on literary history. In a word, the literary scholar hopes to be surprised by the computer scientist. However, surprise is too easy a commodity to supply in data mining. Consider that some of the first literary data miners were the Dadaists and Surrealists, who produced poetry by cutting a printed text nto pieces and pulling those pieces randomly from a bag. In machine learning, this method of textual analysis is known as Gibbs sampling (Duda), and has been used in recent work on probabilistic author-topic modeling (Steyvers). This sort of surprise, however, may not be that which a literary scholar desires--it may not be a meaningful surprise. Thus, the scholar must define for the machine-learning specialist exactly what sort of surprises are desired, so that the appropriate data mining methods may be applied. This is a curious hermeneutic circle--the critic worries that requesting a particular kind of surprise effectively removes true surprise from the process. Data exploration requires a bound on the unknowns to be meaningful and productive. We adduce examples of this need with experiments in anomaly detection on literary data.

Data exploration may be performed by employing data visualization techniques, or by using unsupervised methods of machine learning such as clustering. In both of these situations, it is important to keep the cartographer's dilemma in mind. In order to understand large data sets in high-dimensional space both the literary scholar and the computer scientist require some form of dimensionality reduction. While reductive methods may, indeed, enable new insights, they may also produce artifacts--strange islands analogous to the distorted, massive projection of Greenland on most two-dimensional maps of the world--that give a distorted view of the underlying structure.

Two specific dangers, then, accompany data exploration. The first is that a distorted artifact, a picture, may be mistaken for an underlying truth. The second is that once a data set has been fully explored, it may no longer be valid to use it for hypothesis testing. An exhausted data set prompts us to move on to a new set of texts, to generalize as discussed above. But moving to a new set of data, we often discover that our hypothesis is not portable and fails to generalize. The history of literature is a "collective system," as described by Franco Moretti in Graphs, Maps, and Trees (4), but law-like generalizations are difficult to frame and even harder to transport from text collection to text collection: we discover patterns, yes, and congruent patterns

may be discovered in different collections. To say more is to abandon many of the certainties that the literary scholar had hoped machine learning would provide him with. In preparing a data set we construct a system; leaving that data set behind we leave behind its artificial systematicity as well. Here the literary scholar is surprised to find the computer scientist a more thoroughgoing poststructuralist than himself.

It may seem that data mining offers no more claims to objectivity than literary scholarship -- and indeed, from a certain perspective, this is the case. At its worst, data mining and visualization techniques produce mere inkblots that do little more than manifest the hidden (and indeed, perhaps even unknown) biases of the researchers. However, these explorations gain in consequence as the tacit assumptions of both the literary scholar and the data miner are clearly stated. To assist in foregrounding assumptions, we propose a protocol for researchers in these disparate fields. This protocol includes ways of talking about patterns in a common language, for defining meaningful data representations, and for selecting appropriate statistical assumptions. Only when careful preparatory work is done can data mining have a claim to meaning in the humanities.

Bibliography

Duda, R. O., P. E. Hart, and D. G. Stork. *Pattern Classification.* 2nd Edition. Wiley-Interscience, 2000.

McCarty, Willard. "Modeling: A Study in the Meaning of Words." *A Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Malden, MA: Blackwell Publishing Ltd, 2004. 255-70.

Moretti, Franco. *Graphs, Maps, Trees.* London: Verso, 2005.

Salzberg, Steven. "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach." *Data Mining and Knowledge Discovery* 1.3 (1997): 317-328.

Steyvers, M., P. Smyth, M. Rosen-Zvi, and T. Griffiths. "Probabilistic Author-Topic Models for Information Discovery." *Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining.* New York: ACM Press, 2004. 306-315.

Reading Tools, or Text Analysis Tools as Objects of Interpretation

Stéfan Sinclair *(sgsinclair@gmail.com)*
McMaster University
Geoffrey Rockwell *(georock@mcmaster.ca)*
McMaster University

Computer-assisted text analysis has over 50 years of history in providing tools to help scholars interpret texts (see, for instance, Potter, 1991, Burrows, 2004, Bradley, 2004). The tools themselves are the products of a large range of circumstances in computing and text criticism, from the availability of certain hardware to the fluctuating fortunes of structuralist approaches in literary criticism. In this paper we will reverse the usual interpretive flow from tools to text by attempting to interpret tools themselves as artifacts of human creation. We will frame this interpretive exercise as a Gedankenexperiment; in particular: if a scholar one hundred years from now were to study the TAPoR Portal as cultural artifact, what would it reveal about its theoretical presuppositions, the methodological practices of its times, its cultural value, and even its authors?

This thought experiment is predicated on the assumption that tools can be studied as cultural artifacts in ways similar to, say, literary texts. To examine this assumption more closely, we will outline several ways in which tools may be studied and interpreted, including:

- as functional systems fulfilling an identified need
- as code that corresponds to certain expectations in terms of structure, brevity, creativity, etc.
- as interfaces that may have an aesthetic appeal
- as pedagogic tools that are intended to assist users in developing skills
- as artifacts that express an author's perspective
- as artifacts that express characteristics about a community

Several differences are evident between literary texts and tools as objects of study. For instance, tools manifest themselves at two (at leaset) layers of visibility: the code layer, generally reserved for developers, and the interface layer, generally intended for users. Literary texts, in contrast, have only one layer of exposure. It may also be that text analysis tools are

fundamentally too different from texts to be considered using approaches of literary criticism. A similar debate has raged in game studies for the past several years (see, for instance Frasca, 1999): can games be studied as narratives (the literary camp) or do they require an entirely different approach (the ludology camp)? We will argue for a hybrid approach: while literary criticism can be useful to interpret certain aspects of tools, other aspects require their own framework of study. Moreover, this hybrid approach leads to certain practical consequences when considering how to peer-review tools and their development (see Sinclair et al., 2003). As a case study for interpreting tools, we will use the TAPoR Portal, an initiative to build a web-based gateway to electronic texts and tools (see <http://tapor.ca/> and Sinclair, 2002). The TAPoR Portal is reaching the end of its development cycle, after approximately five years. It serves as a convenient case study for several reasons, including its duration (a five year project), its collaborative nature (involving several dozen researchers at five Canadian universities, its development model (a blend of academic and private sector contributions), its interdisciplinarity (representing several branches of the digital humanities and computer science), and, of course, its size and complexity (over 165, 000 new lines of code). Indeed, given the size of this "corpus", examples can be found for just about any interpretation that might be proposed, but such is the open-ended nature of interpretation.

Bibliography

Bradley, John. "Text Tools." *A Companion to Digital Humanities*. Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing Ltd, 2004. 505-522.

Burrows, John. "Textual Analysis." *A Companion to Digital Humanities*. Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing Ltd, 2004. 323-347.

Frasca, Gonzalo. "Ludology Meets Narratology: Similitudes and Differences Between (Video) Games and Narrative." 1999. Accessed 2006-11-13. <http://www.ludology.org/articles/ludology.htm>

Potter, Rosanne G. "Statistical Analysis of Literature: A Retrospective on Computers and the Humanities, 1966-1990." *Computers and the Humanities* 25.6 (1991): 401-429.

Sinclair, Stéfan, John Bradley, Stephan Ramsay, Geoffrey Rockwell, and Ray Siemens. "Peer Review of Humanities Computing Software." *ACH/ALLC 2003 Conference Program.* 2003. Accessed 2006-11-13. <http://www.english.uga.edu/webx/abstracts/final/sessions.pdf>

Sinclair, Stéfan, and Terry Butler. "TAPoR - A Canadian Text Analysis Portal for Research." *Digital Resources in the Humanities*. London: Office for Humanities Communications, 2002.

TAPoR Project. Accessed 2006-11-13. <http://www.tapor.ca/>

Agora.Techno.Phobia.Philia: Gender, Knowledge Building, and Digital Media

Martha Nell Smith *(mnsmith@umd.edu)*
University of Maryland (MITH)
Carolyn Guertin *(carolyn.guertin@gmail.com)*
University of Texas at Arlington
McLuhan Program in Culture and Technology
Laura C. Mandell *(mandellc@muohio.edu)*
Department of English
Miami University
Katherine D. Harris *(kharris@email.sjsu.edu)*
San Jose State University

The editors of a special issue of *Signs: Journal of Women in Culture and Society*, a prominent scholarly feminist journal, wrote in 1990 that "the degree to which American society has embraced and absorbed computer technologies is astonishing. The degree to which the changes provoked by computers leave prevailing inequalities is troubling."[1] This observation preceded the development of the World Wide Web, which has enabled computational tools to suffuse much work of the humanities. The questions that have informed our work as feminist theorists and scholars—how do our items of knowledge come into being, who made them, for what purposes, and how does gender play a role in knowledge making—inhere in our digital humanities work. That the two fields are or should be inextricably intertwined seems, therefore, an inevitable fact of life. But is this just personal coincidence, a fact produced by the trajectory of our careers and interests? What is humanities computing anyway, and why should it be important for feminist cultural, social, and intellectual work? Concomitantly, can feminism enhance and improve the world and work of computer science, of humanities computing, of digital humanities? After all, "very early in life, computing is claimed as a male territory. At each step from early childhood through college, computing is both actively claimed as "guy stuff" by boys and men and passively ceded by girls and women. The claiming is largely the work of a culture and society that links interest and success with computers to boys and men."[2]

A culture that says to use computing tools expertly one must know how machines work, or at least must be deft programmers, dominates much of the world of humanities computing. It is as if those who have fretted over literary and other humanities fields becoming feminized or soft have been rescued by a field that is hard science. Thus through computing, humanities is being remasculinized. Scientific matters of mathematics and computation, objective and hard, are not subject to the concerns of gender, race, or sexuality. Either explicitly or implicitly, concerns that had taken over so much academic work in literature—of gender, race, class, sexuality—were assumed to be irrelevant to humanities computing. 2 + 2, so the reasoning goes, always equals 4, whether you are black, female, queer, or straight. The codes always work, whatever one's personal identity or social group, and, as matters of objective and hard science, are best dealt with by those who have been most interested in being engineers and computational scientists of critical inquiry. So surely those interested are also folks who do not want to clutter sharp, disciplined, methodical philosophy with considerations of the gender-, race-, and class-determined facts of life. After all, in the wake of the sixties, the humanities in general and their standings in particular had suffered, according to some, from being feminized by these things. Humanities computing seemed to offer a space free from all this messiness and a return to objective questions of representation.

Yet such dreams of a return to the "objective," uncluttered by messy identity questions, are nostalgic. That humanities disciplines were in fact foundationally changed by feminist scholarship of the twentieth century is obvious from project development within humanities computing itself, and each of the panelists is deeply involved with a major digital humanities project informed by feminist scholarship—the *Dickinson Electronic Archives*, *The Poetess Archive*, *The Forget-Me-Not Hypertextual Archive*, a *cyberfeminist archive*, *The Assemblage Gallery*. Our work as feminists leads us to concur with social scientist Jane Margolis and computer scientist Allan Fisher that "the goal is not to fit women into computer science [digital humanities] as it is currently taught and conceived. Rather, a cultural and curricular revolution is required to change computer science [digital humanities] so that the valuable contributions and perspectives of women are respected within the discipline."[3] In other words, this panel will posit ways in which the methods of feminist, queer, and race critical inquiry might benefit the work of digital humanists across the board (or screen, as it were) rather than serving as a special niche of interest (for example, the methods might benefit the multi-institutional NORA project, <http://noraproject.org>, and its second phase, MONK, tremendously).

Descriptions of the roundtable presenters are below and we trust will serve to show the range of experience and expertise on which we will draw to pose our questions, posit ways in

which the messiness of such critical inquiry can advance digital humanities, entertain questions and suggestions before, during, and after our session in order to collaborate with audience members in knowledge production.

Martha Nell Smith is Professor of English and Founding Director of the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland. Author of more than 40 articles, including "Electronic Scholarly Editing" in the *Blackwell Companion to Digital Humanities*, her publications include *Open Me Carefully: Emily Dickinson's Intimate Letters to Susan Dickinson*, with Ellen Louise Hart (1998); *Comic Power in Emily Dickinson*, with Cristanne Miller and Suzanne Juhasz (1993); and *Rowing in Eden: Rereading Emily Dickinson* (1992). With Mary Loeffelholz, she is editing the *Blackwell Companion to Emily Dickinson* (forthcoming in 2007). She is also Coordinator and Executive Editor of the *Dickinson Electronic Archives* projects at the Institute for Advanced Technology in the Humanities (IATH) at the University of Virginia. With Lara Vetter, she is a general editor of *Emily Dickinson's Correspondence: A Born-Digital Inquiry*, forthcoming from the Mellon-sponsored University of Virginia Press Rotunda Electronic Imprint. Her digital humanities work is an extension of her work as a feminist literary theorist and scholar. Because her interest in the possibilities afforded by computers as powerful and empowering tools of humanities scholarly work became so keen as the World Wide Web was gaining precedence, her work in humanities computing has been powerfully influenced by cyberculture and new media studies. As a digital humanities specialist she has focused on the sociologies of knowledge production in our technology-saturated world—what data is reproduced and made accessible, and to whom, and what new knowledge has been produced by computational tools. Her questions, suggestions, and models will be drawn from the multi-institutional data mining and visualization NORA Project (<http://noraproject.org>).

Carolyn Guertin is an Assistant Professor of Digital Media and Director of the eCreate Lab in the Department of English at the University of Texas at Arlington. During the 2004 to 2006 academic years, she was a Senior McLuhan Fellow and SSHRC Postdoctoral Fellow in the McLuhan Program in Culture and Technology at the University of Toronto ~ most recently giving the closing keynote address at "Re-Reading McLuhan: An International Conference on Media and Culture in the 21st Century" at the University of Bayreuth in Germany. She does both theoretical and applied work in cyberfeminism, digital narrative, digital design, media literacy (or postliteracy) and performance. She is a founding editor of the online journal *MediaTropes*, and a literary advisor to the Electronic Literature Organization. She has written textbooks on hypertext and literature and information aesthetics, and is currently working on a new book project called *Connective Tissue: Queer Bodies, Postdramatic Performance and New Media Aesthetics*. Guertin is best known as curator and founder of *Assemblage: The Women's New Media Gallery* <http://tracearchive.ntu.ac.uk:80/traced/guertin/assemblage.htm>, the only site devoted exclusively to born-digital art and lit by women on the Web ~ soon to be relaunched in a 2.0 version. She will be examining the trend toward personal media and participatory culture that is a product (or fallout) of the politicizing of diversity, the women's movement, and queer issues. Where notions of interactivity focused on the technology as the most important component, participatory or user-generated culture—from wikis to podcasts to FaceBook to *Second Life*—puts the backchannel into the foreground, and puts people of all genders back as active users in the system.

Laura Mandell is Associate Professor in eighteenth-century and Romantic British literature at Miami University. Her book, *Misogynous Economies: The Business of Literature in Eighteenth-Century Britain*, discusses the feminist potential of anti-feminist writings produced during the long eighteenth century. She has published essays in ELH, MLQ, European Romantic Review, Studies in Romanticism, and Nineteenth-Century Prose. She has edited The Castle of Otranto and *The Poetess Archive Database* (<http://unixgen.muohio.edu/~poetess>), a TEI-encoded bibliographic finding aid and full-text resource about the men and women who wrote popular poetry in Britain and America between 1750 and 1900. Within the next year, this resource will contain author and title information from tables of contents of all the major anthologies and literary annuals. Within the next few years, it will expand to periodicals.

Katherine D. Harris, Assistant Professor, San Jose State University, has created an online hypertextual archive of the first literary annual, the Forget Me Not, re-presenting various aspects of the book as well as the poetry, prose and engravings: "Forget Me Not: A Hypertextual Archive of Ackermann's 19th-Century Literary Annual" <http://www.orgs.muohio.edu/anthologies/FMN/Index.htm>. With Laura Mandell, she serves as an editor of *The Poetess Archive Database*, which now contains a bibliography of over 4,000 entries for works by and about writers working in and against the "poetess tradition," the extraordinarily popular, but much criticized, flowery poetry written in Britain and America between 1750 and 1900. Their presentation for this panel will not be a show-and-tell of these archives, but an in-depth consideration of ways in which the feminist theories that have identified the scholarly needs for this resource and informed their development can advance the work of digital humanities at large.

1. Jean F. O'Barr, ed., *From Hard Drive to Software: Gender, Computers, and Difference*. Special Issue of *Signs: Journal of Women in Culture and Society* 16:1 (Autumn 1990).
2. Jane Margolis and Allan Fisher. *Unlocking the Clubhouse: Women in Computing* (Cambridge, Massachusetts and London England: The MIT Press, 2002), p. 4.
3. *Unlocking the Clubhouse*, p. 6.

Lost in the Archives, Found in Digital Collections

Natalia (Natasha) Smith *(nsmith@email.unc.edu)*
University of North Carolina at Chapel Hill Library
Xie Dongqing *(dongqing.xie@gmail.com)*
University of North Carolina at Chapel Hill Library
Elizabeth McAulay *(elizabethmcaulay1@yahoo.com)*
University of North Carolina at Chapel Hill Library
Todd Cooper *(cojere@email.unc.edu)*
University of North Carolina at Chapel Hill Library
Adrienne M. MacKay *(awmackay@gmail.com)*
University of North Carolina at Chapel Hill Library

Abstract

The proliferation of digital collections on the web has dramatically expanded access to content that would remain otherwise underutilized or undiscovered. Organizations that have been actively involved in large scale digitization—through Google, the Open Content Alliance (OCA), or other means—most often limit the scope of these projects to digital replication of materials. Not surprisingly, given the size and extent of many collections selected for digitization, the costs of applying extensive research and scholarship to primary documents are often prohibitive, if such actions are even considered by project staff. Digital scholarly editions offer many research advantages that exceed the limitations of traditional and linear print publications. Their potential has been already instantiated by a few prominent projects, among them Rotunda by the University of Virginia Press and the Perseus Digital Library. [1] Encouraged by these exemplars, *Documenting the American South* (DocSouth), UNC-Chapel Hill Library's digital publishing program, sought to create two online scholarly documentary histories: a collection of documents related to antebellum student life at

the University of North Carolina and a collection of documents about the institutional development of the university during the same period. Through careful planning and analysis, collaboration with research scholars and subject librarians, and the application of open-source technology and international standards, DocSouth's "True and Candid Compositions: The Lives and Writings of Antebellum Students at the University of North Carolina" and "The First Century of the First State University" [2] represent rare examples of scholarly publications with annotations and interpretive essays that include both color facsimile and transcription access to unique primary documents.

Documenting the American South has ten years of experience encoding printed materials using TEI guidelines; however, creating online documentary histories required a more complex approach than had previously been employed with our collections. "True and Candid Compositions", for example, was originally conceived as a monograph including diplomatic transcription of manuscript documents; textual, biographical, and interpretive annotations; several scholarly essays; and an extensive index—all prepared by Dr. Erika Lindemann, professor of composition at the UNC-Chapel Hill. In presenting Lindemann's project on the web, we strove to include all of the features of her project, plus features only possible through and valuable for an online publication. [See Figure 1 and Figure 2 for screenshots.]

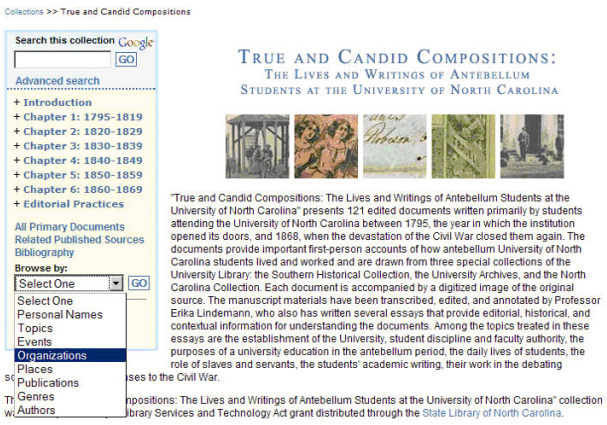


Figure 1: Screenshot of True and Candid index page

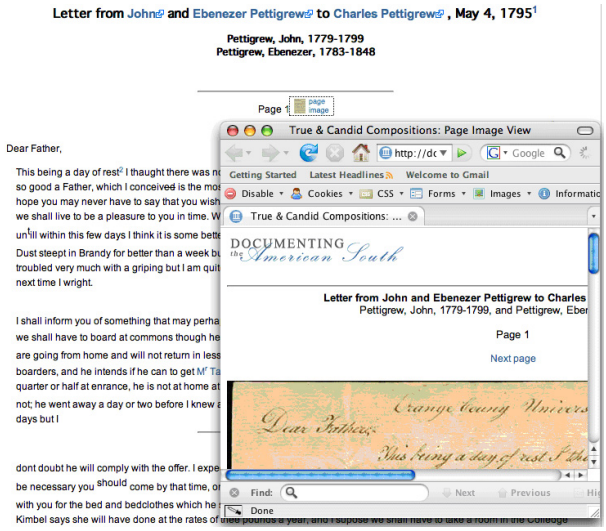


Figure 2: Screenshot of document with image of manuscript

Application of Technologies

In both online publications, we wanted to provide users with several options to fully explore and discover content in these unique collections. We offer several "browse by" indices that were compiled by extracting information from the TEI-XML files. For "True and Candid Compositions" and for "The First Century," a total of 500 transcribed manuscript documents and over 20 scholarly essays were encoded by graduate students from the English Department and School of Information and Library Science using "TEI in the Libraries" recommendations for level 5 of encoding.[3] All personal, place, and organization names were disambiguated by scholars and assigned a unique id number, a regularized name, and one of three type descriptors (i.e., person, place, organization). This information was then encoded within `<name>` elements with relevant attributes as part of each document XML file. Images of all manuscript pages were scanned and saved in TIFF and JPEG formats.

The XML and image files became input for a publishing mechanism comprised of two distinct parts: (1) conversion of XML to XHTML and (2) generation of search and browse functionality in XHTML pages on the web. [See Figure 3: Document Processing Workflow.]

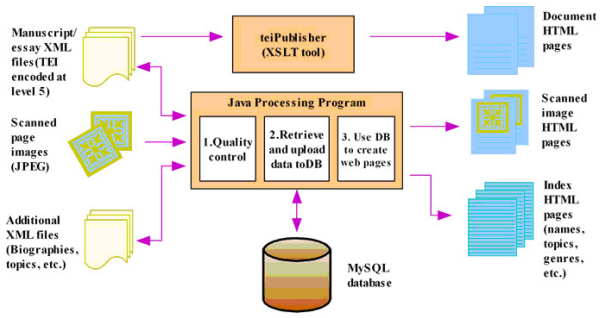


Figure 3: Document Processing Workflow

First, we extracted metadata from each XML file (one XML file per document) and then converted the XML files to XHTML. The resulting XHTML files have the distinctive DocSouth format and include links to images and biographical annotations, both of which open in separate windows (pop-ups). This step is accomplished with the use of the XSLT technology, available from the <teiPublisher> software[4], which we modified to meet our project needs. We used <teiPublisher> with locally customized XSLT files to transform manuscript and essay XML files into XHTML files. Our customizations included: displaying the TEIHeader under tabbed buttons; highlighting names for which regularized names are viewable as a mouseover; adding icons to personal names to indicate that biographical essays are available in pop-up window, and adding URL links to scanned page images in separate XHTML files.

Second, we generated indices with a custom Java program and our primary MySQL database, using JAXP/JDBC/XPath technologies. "XML is a good match for Java. It pairs Java's code portability feature with its data portability…"[5] The Java programming language has a number of proven APIs for working with XML, including the Java API for XML Processing (JAXP). In concert with XPath and JDBC (Java API for database interactions), the language is a powerful tool for performing a variety of operations with XML documents. Additionally, Java is platform agnostic and secure. We designed this program to perform a wide range of operations, including: reporting errors in XML files for quality control; retrieving a variety of data from XML files; exporting data into the MySQL database; generating an XHTML file for every page image of each document; and extracting data from the database to create indices of personal names, geographic names, organization names, authors, genres, and topics.

This Java program extracted rich metadata from the XML files, specifically from the `<teiHeader>` and from the key, reg, and type attributes of the `<name>` element. These varied data were stored in the primary DocSouth MySQL database, which contains all the metadata for all DocSouth online collections. In processing, the Java program extracted all encoded names and recorded their regularized names, types, and occurrence locations in the XML files to the database. In a similar process, biographical information contained in an additional XML file was processed and stored in the same database. A number of scholars and subject librarians wrote this expansive biographical information for hundreds of identified and researched proper names, which has served to add valuable contextual richness to the collection.

We also designed our Java program to generate two types of "browse by" indices—proper name indices and document indices. The program generated proper name indices by type; these indices are hyperlinked to a PHP program that searches the database and provides a list of documents in which the name appears. Each linked document displays a document page highlighting the selected name. Finally, the program uses available metadata to generate document indices that offer other chapter-level "browse by" options, including: documents listed by chapter, topic, genre, and author. As a result of this Java processing, these two online documentary histories have 2,340 personal names, 170 organization names, and 420 place names in indices generated from the XML files.

DocSouth's "True and Candid Compositions" and "First Century" collections represent exciting additions to our digital library. More than just collections of digitized manuscripts, these online histories benefit from collaboration with scholars and DocSouth's considerable experience and technical expertise. The confluence of unique primary documents, scholarship, and judicious application of open-source technology solutions allowed us to develop two online scholarly documentary histories that far exceed the limitations of traditional print publications. No longer lost in the archives, these enhanced digitized primary source materials are readily available to enthusiasts, scholars, and learners everywhere.

---

1. `<http://rotunda.upress.virginia.edu/index.php?page_id=About >` and `<http://www.perseus.tufts.edu/>`
2. "True and Candid Compositions: The Lives and Writings of Antebellum Students of the University of North Carolina." *Documenting the American South* Ed. Erika Lindeman. (University of North Carolina at Chapel Hill Libraries). `<http://docsouth.unc.edu/true/>`"The Carolina at Chapel Hill Libraries
3. The TEI Consortium, "TEI Text Encoding in Libraries. Guidelines for Best Encoding Practices" `<http://www.diglib.org/standards/tei.htm>`.
4. `<http://sourceforge.net/projects/teipublisher/>`
5. Akmal B. Chaudhri, *XML Data Management: Native XML and XML-Enabled Database Systems* (Boston: Addison-Wesley, 2003) p. 342.

Digital Innovations in Teaching and Learning: Interactive Computer Environments in the Undergraduate Classroom

Lisa M. Snyder *(lms@ats.ucla.edu)*
University of California, Los Angeles

Instructional use of interactive three-dimensional computer models is transforming undergraduate education at the University of California, Los Angeles. A surge of faculty interest in virtual environments over the past ten years has resulted in a broad spectrum of projects now making their way into Humanities and Social Sciences classrooms. Art history and architecture students can now interactively explore the digital Roman Forum developed by UCLA's Experiential Technologies Center (the successor organization to the Cultural Virtual Reality Lab) in place of filmic slides or PowerPoint presentations. Near Eastern Languages and Cultures students can tour through Qumran, the settlement associated with the Dead Sea Scrolls, in a computer reconstruction developed by the Qumran Virtual Reality Project, or compare the first century Herodian Temple Mount with the eighth century Umayyad structures on the site through the real-time visual simulation model developed jointly by the Urban Simulation Team at UCLA and the Israel Antiquities Authority. Archaeology students can experience the ancient Egyptian sites of Karanis and Karnak created with support from UCLA's Office of Instructional Development and Academic Technology Services. Spanish and Portuguese students can culminate their studies of the pilgrimage route in Spain with a virtual visit to the Romanesque Cathedral of Santiago de Compostela, complete with authentic period music. In American History classrooms, students can experience the wonders of the World's Columbian Exposition of 1893 by interacting with a model developed by the Urban Simulation Team.

The proposed paper will describe the results of over 600 student surveys administered in the past two years by the UCLA Experiential Technologies Center (ETC) staff to solicit reactions to this new form of instructional technology. The survey instruments were completed following regularly scheduled class meetings held either in a technology-enabled classroom or UCLA's Visualization Portal (a campus facility with a 160 degree spherically wrapped projection screen specifically designed for displaying these virtual environments). In the surveys, Likert-style ratings gauged the students' overall experience with the computer model, their understanding and interest in the content of the virtual environment, and their reactions to the technology as a learning tool and compared to more traditional types of instructional technologies. Multiple choice and ranking questions explored the students' interest in using the virtual environments outside of the classroom and the aspects of the environment most important for creating an engaging learning experience. Short answer questions delved into the students' likes and dislikes, and thoughts on the learning benefits of interactive computer models.

The paper will also explore instructor reactions to the classroom use of interactive computer models. Concurrent with the student surveys, ETC staff administered instructional technology questionnaires to twenty five undergraduate instructors from around the country who participated in an NEH Summer Institute focused on "Models of Ancient Rome" and conducted personal interviews with twelve instructors actively using the models in their classrooms to explore faculty reactions to teaching with virtual environments. The results of these surveys identify the perceived challenges and benefits to classroom use of interactive computer environments, general concerns about instructional technology, curricular integration, perceived and experienced pedagogical impacts, and instructor expectations for virtual environments.

The paper will conclude with an analysis of how the student and instructor reactions to the UCLA environments are informing ongoing project development, and a discussion of future research regarding digital pedagogy in the Humanities.

Associated websites:

- UCLA Experiential Technologies Center (<http://www.etc.ucla.edu>)
- UCLA Academic Technology Services (<http://www.ats.ucla.edu>)
- The Urban Simulation Team at UCLA (<http://www.ust.ucla.edu>)
- UCLA Office of Instructional Development (<http://www.oid.ucla.edu>)

Scholarly (R)evolution: Roles of E-texts in the Research Process in the Humanities

Suzana Sukovic *(suzana.sukovic@uts.edu.au)*
University of Technology, Sydney

Introduction

The study on the roles of electronic texts (e-texts) in the research process in the humanities investigates how academics in literary and historical studies work with electronic textual resources and how interaction with electronic texts affects their research processes. It is situated in the context of discussions about the cultural revolution triggered by the development of digital media. The common comparison with the print revolution suggests that textual sources are at the centre of the change, which is particularly significant for scholars in the humanities. Since text is the basic working material for research in the humanities, the provision of text in an adequate electronic form is crucial for qualitatively new applications of computer technology in humanistic disciplines. At the same time, an understanding of scholars' interactions with e-texts is critical for decisions on how to present existing sources and for any work on further developments.

In his seminal book, *Radiant Textuality*, Jerome McGann talks about the 'material revolution' in which we reconceive the entity of our cultural archive of materials. Since these repositories provide the basis for all traditional scholarly work, institutional changes have been having a radical effect on the traditional scholarship (McGann 2001). The acceptance of digital resources in the humanities has not been as clear and as decisive as in sciences, but reports from different countries over the last few years point to a positive shift in the quantity and quality of scholars' engagement with information and communication technologies (Houghton, Steele, and Henty 2003; The British Academy 2005; American Council of Learned Societies 2006)

A range of studies explored the information behaviour of the humanities scholar; some were focused on the use of digital resources in the humanities but none dealt primarily with the use of electronic texts. Several studies have been conducted, dealing briefly with the use of electronic texts as part of a broader investigation (Massey-Burzio 1999; Brockman et al. 2001) or surveying the use of a particular electronic resource (Flanders 1998; Porter 1998; Duff and Cherry 2000; Cherry and Duff 2002). The literature suggests that electronic texts, while much appreciated by those who use them, have not become widely accepted, even in disciplines that are heavily based on textual studies (Warwick 1999; Brockman et al. 2001). Studies of citation patterns, such as Graham's investigation of citations in historians' professional publications, show that electronic resources do not rate highly in published works (Graham 2000, 2001). However, citation patterns may indicate intellectual exchanges to some extent, but they are not an accurate reflection of the use of electronic resources. Palmer (2005) points out that there are scholars who have begun to create digital resources for themselves, which is one of the indications of how scholars wish to engage new technologies in their research.

There is a gap in our knowledge about the roles of electronic texts in the research process so we need to explore various aspects of scholars' interactions with e-texts and explain how these interactions contribute to the research process.

Methodology

The exploratory study on the roles of electronic text in the research enquiry used qualitative methodology to investigate research projects in which e-texts have been used and the nature of academics' interactions with these texts. The following research questions guided the development of the study:

1. How do academic researchers in literary and historical studies work with electronic texts?
    - How and for what purposes do researchers interact with electronic texts?
    - How do researchers think and feel about the research context in which they work with e-texts?
2. How are the interactions with e-texts integrated in the research process?
3. What is the contribution of e-texts to the research process?
4. What are the obstacles and aids in engagement with electronic texts?

The study has dealt with the use of electronic texts as a resource and tool, as opposed to projects that aim to produce electronic textual editions or enhance electronic texts in any way. The participants saw their work as traditional humanities research or discussed their projects aiming to have traditional outputs. Investigated research projects were in the areas of literary and historical studies, because both fields are known for extensive and sophisticated use of textual resources. Participants from six universities in two Australian cities and one participant from

a university in the USA (altogether 16 participants) participated in the study and discussed thirty research projects.

The term 'electronic text' in this study means any textual material in electronic form, used as a primary source in literary and historical studies. Primary materials are usually poetry, stories, novels, plays, and a variety of historical documents — government, public or private. Digitised archival copies of magazines and newspapers, as well as web-sites and blogs, could be electronic texts as defined here when they are used as primary sources. Electronic texts could be written or spoken (e.g., oral histories), digitised or created electronically, stand-alone documents or part of electronic databases and editions.

The study has had two phases. The first phase included in-depth semi-structured interviews, examination of participants' manuscripts and published works as well as examination of some e-texts they mentioned during interviews. The second phase involved detailed data-gathering from a small group of academics drawn from the participants in the first phase. The grounded theory techniques described by Strauss (1987), Strauss and Corbin (1998a; 1998b) and Glaser (1998) were used for data analysis.

Roles of e-texts

This paper presents findings directly related to roles of e-texts in the research process but the understanding of the roles is based on other findings that emerged from the study. Firstly, scholars in the study perceived e-texts as fluid entities, which combine different media and formats in a way that does not match the traditional divisions of library materials. The perceived fluidity of electronic textuality leads to converging and transformed practices of networking and information searching. These practices combine aspects of networking, chaining, browsing and web-surfing in traditional and new ways so that the pattern of a new information behaviour emerges. I called this new practice *netchaining* (Sukovic 2006).

The roles of e-texts are based to a large extent on working with e-texts as fluid entities. Four main roles of e-texts in the research process emerged from the study.

1. Support in finding documents and information. Search capabilities combined with the provision of full text documents provide a powerful aid in information discovery. E-texts provide support in information retrieval and discovery of primary materials; they lead to other sources and aid in working with analogue sources; supplement hard copies and contribute to the current awareness.

This is the most fundamental role. Not only do information discovery and retrieval provide a basis for all other roles, but the nature of scholars' interaction with e-texts during the retrieval determines other roles to a large extent (Role 3, for example).

2. Aid in managing the research process. Access to e-texts allows scholars to plan visits to remote collections; aids the publication process and provides sources for some research activities (e.g., ordering digital images, confirming publication rights, and exchanging files with collaborators and publishers).

3. Aid in investigation of the topic. The multiplicity of sources, formats and textual information that could be quickly brought together is a basis of exploration that allows scholars to see different meanings and aspects of the topic. Exploration of research questions through interactions with e-texts took four main forms:

- exploration of patterns and connections by searching and comparing diverse bodies of electronic texts;
- production and/or interrogation of textual databases to explore research questions;
- exploration of electronically born literature and
- exploration as part of the academic research to be used in creative.

4. Contribution to writing and presenting research results. The use of e-texts improves the speed and accuracy in writing by allowing copying and pasting of passages. Interactions with e-texts and digital media in general promote new, less structured and linear ways of thinking about the topic, which influences the academic writing style. From subtle changes in presenting the argument to more radical combinations of academic and creative writing styles, the participants reflected on different ways in which electronic textuality was influencing their traditional academic writing. Interactions with e-texts also encourage thinking about alternative modes for presenting research findings that do not fit traditional academic genres. E-texts contribute to the final research stages in a complex process of negotiation with the research tradition.

These roles serve two main functions:

- aid in providing basis for research (support roles) or
- aid in exploring the topic and presenting research findings (substantive roles).

E-texts provide basis for research, or play support roles, when they make some aspects of the research process quicker and easier. The speed and convenience, or frustration sometimes associated with working with electronic sources, may influence the process, but they normally do not affect the scholar's intellectual engagement with the topic in a significant way. The first two roles, Support in finding information and Aid in managing research process, are support roles. The fourth, Contribution to writing and presenting research results, plays

a support role when e-texts help in improving the speed and accuracy.

E-texts aid in exploring the topic and presenting research findings, or play substantive roles, when they take part in shaping the scholar's thinking process. Interactivity is an essential element in following hunches, testing hypotheses and making connections in a way that was impossible or impractical without e-texts. The scholar's thinking about the topic develops in the interplay with the e-text and this experience can influence the presentation of research results. The third role, Aid in investigation of the topic, is a substantive content-oriented role. In Contribution to writing and presenting research results, e-texts have a substantive role when they influence the writing style and presentation of research results.

Conclusion

Understanding of the roles of e-texts in the research process contributes to our understanding of the envisaged scholarly change as well as information needs and behaviour in the humanities. It confirms the recent reports on the change of research practices resulting from the use of ICTs and explores the impacts of interactions with e-texts on the research process. The study can have practical implications for the development of digital collections and software applications, approaches to text encoding and development of training programs and institutional policies.

Bibliography

American Council of Learned Societies' Commission on Cyberinfrastructure for Humanities and Social Sciences. *Our Cultural Commonwealth: the report of the American Council of Learned Societies' Commission on Cyberinfrastructure for Humanities and Social Sciences.* 2006. Accessed 2006-08-12. <http://www.acls.org/cyberinfrastructure/acls.ci.report.pdf>

British Academy. *E-resources for Research in the Humanities and Social Sciences: A British Academy Policy Review*. London: The British Academy, 2005.

Brockman, William S., Laura Neumann, Carole L. Palmer, and Tonyia J. Tidline. *Scholarly Work in the Humanities and the Evolving Information Environment.* Washington, D.C.: Digital Library Federation and Council on Library and Information Resources, 2001. Accessed 2006-06-30. <http://www.clir.org>

Cherry, Joan M., and Wendy M. Duff. "Studying Digital Library Users Over Time: A Follow-up Survey of Early Canadiana Online." *Information Research* 7.2 (2002). Accessed 2002-06-06. <http://informationr.net/ir/7-2/paper123.html>

Duff, Wendy M., and Joan M. Cherry. "Use of Historical Documents in a Digital World: Comparisons with Original Materials and Microfiche." *Information Research* 6.1 (2000). Accessed 2004-05-20. <http://informationr.net/ir/6-1/paper86.html>

Flanders, Julia. " Scholarly Research and Electronic Resources." *WWP Newsletter* 4.2 (1998). Accessed 2004-05-28. <http://www.wwp.brown.edu/project/newsletter/vol04num02/scholarly042.html>

Glaser, Barney G. *Doing Grounded Theory: Issues and Discussions.* Mill Valley, CA: Sociology Press, 1998.

Graham, Suzanne R. "Historians and Electronic Resources: A Citation Analysis." *Journal of the Association for History and Computing* 3.3 (2000).

Graham, Suzanne R. "Historians and Electronic Resources: A Second Citation Analysis." *Journal of the Association for History and Computing* 4.2 (2001).

Houghton, John W., Colin Steele, and Margaret Henty. *Changing Research Practices in the Digital Information and Communication Environment.* [Canberra?]: Department of Education, Science and Training, 2003. Accessed 2005-05-11. <http://www.dest.gov.au/sectors/research_sector/publications_resources/profiles/changing_research_practices.htm>

Massey-Burzio, Virginia. "The Rush to Technology: A View from the Humanities." *Library Trends* 47.4 (1999): 620–639.

McGann, Jerome. *Radiant Textuality: Literature After the World Wide Web.* New York: Palgrave, 2001.

Palmer, Carole L. "Scholarly Work and the Shaping of Digital Access." *Journal of the American Society for Information Science and Technology* 56.11 (2005): 1140-1153.

Porter, Sarah. "Reports from the Front: Six Perspectives on Scholar's Information Requirements in the Digital Age." *The New Review of Academic Librarianship* 4 (1998): 167–189.

Strauss, Anselm. *Qualitative Analysis for Social Scientists.* Cambridge, UK: Cambridge University Press, 1987.

Strauss, Anselm, and Juliet Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 1987. Second Edition. Thousand Oaks, CA: SAGE Publications, 1998.

Strauss, Anselm, and Juliet Corbin. "Grounded Theory Methodology: An Overview." *Strategies of Qualitative Inquiry*. Ed. N. K. Denzin and Y. S. Lincoln. Thousand Oaks, CA: SAGE Publications, 1998.

Sukovic, Suzana. " Electronic Texts in the Humanities: Converging Media, Formats, Practices and Ideas.." Paper read at IR 7.0: Internet Convergences, presented by the Association of Internet Researchers (AoIR), at Brisbane, 27–30 September. 2006. Accessed 2006-10-24. <http://conferences.aoir.org/index.php?cf=5>

Warwick, Claire. "English Literature, Electronic Text and Computer Analysis: An Impossible Combination? ." Paper read at ACH-ALLC '99 International Humanities Computing Conference, June 9-13, at Charlottesville, Virginia. 1999. Accessed 2002-04-05. <http://InformationR.net/ir/5-2/paper71.html>

A Statistical Study of Superlatives in Dickens and Smollett: A Case Study in Corpus Stylistics

Tomoji Tabata *(tabata@lang.osaka-u.ac.jp)*
University of Osaka

1. Introduction

This study gives a quantitative overview of the use of superlatives in Dickens in comparison with Smollett. The focus is laid on the differing distribution of superlatives in the texts written by the two authors. By applying correspondence analysis, this study tries to illustrate how sharply the two authors differ in their uses of superlatives as well as how texts are clustered according to chronology within authorial sets.

Despite a number of studies on Dickens' style have noted a tendency for overstatement in his fiction (Brook 1970; Sorensen 1985; Golding 1985; Hori 2004, etc.), surprisingly little attention has been paid to superlatives as a whole. Apart from Dickens studies, however, Biber et al. (1999) gives an interesting account of superlatives in four linguistic registers: conversation, fiction, news, and academic prose. According to Biber et al., —*est* superlative adjectives are most frequent in news reportage (c. 1400 times per million words) while "the comparatively low frequency of superlatives in academic writing (c. 800 per million) reflects a general reluctance to make extreme claims" (Biber et al., 521), with fiction showing even lower frequency for the word class (c. 700 per million).

No.	Author	Texts	Abbr. Tags	Category	Date	Word-tokens
1	Dickens	*Sketches by Boz*	(SB)	Sketches	1833-6	188,591
2	Dickens	*The Pickwick Papers*	(PP)	Serial Fiction	1836-7	303,182
3	Dickens	Other Early Papers	(OEP)	Sketches	1837-40	67,149
4	Dickens	*Oliver Twist*	(OT)	Serial Fiction	1837-9	159,256
5	Dickens	*Nicholas Nickleby*	(NN)	Serial Fiction	1838-9	325,345
6	Dickens	*Master Humphrey's Clock*	(MHC)	Miscellany	1840-1	46,087
7	Dickens	*The Old Curiosity Shop*	(OCS)	Serial Fiction	1840-1	219,558
8	Dickens	*Barnaby Rudge*	(BR)	Serial Fiction	1841	256,082
9	Dickens	*American Notes*	(AN)	Sketches	1842	102,068
10	Dickens	*Martin Chuzzlewit*	(MC)	Serial Fiction	1843-4	339,906
11	Dickens	*Pictures from Italy*	(PFI)	Sketches	1846	72,636
12	Dickens	*Dombey and Son*	(DS)	Serial Fiction	1846-8	344,851
13	Dickens	*David Copperfield*	(DC)	Serial Fiction	1849-50	358,720
14	Dickens	*A Child's History of England*	(CHE)	History	1851-3	163,188
15	Dickens	*Bleak House*	(BH)	Serial Fiction	1852-3	357,048
16	Dickens	*Hard Times*	(HT)	Serial Fiction	1854	104,322
17	Dickens	*Little Dorrit*	(LD)	Serial Fiction	1855-7	340,657
18	Dickens	*Reprinted Pieces*	(RPC)	Sketches	1850-6	92,091
19	Dickens	*A Tale of Two Cities*	(TTC)	Serial Fiction	1559	136,625
20	Dickens	*The Uncommercial Traveller*	(UT)	Sketches	1860-9	143,148
21	Dickens	*The Great Expectations*	(GE)	Serial Fiction	1860-1	186,248
22	Dickens	*Our Mutual Friend*	(OMF)	Serial Fiction	1864-5	328,961
23	Dickens	*The Mystery of Edwin Drood*	(ED)	Serial Fiction	1870	94,642
					Sum of word-tokens in the set of Dickens texts:	**4,730,361**
24	Smollett	*Roderick Random*	(SRR)	Fiction	1748	192,910
25	Smollett	*Peregrine Pickle*	(SPP)	Fiction	1751	342,963
26	Smollett	*Ferdinand Count Fathom*	(FCF)	Fiction	1753	159,088
27	Smollett	*Sir Launcelot Greaves*	(SLG)	Fiction	1760	89,636
28	Smollett	*Travels through France and Italy*	(TFL)	Sketches	1766	121,168
29	Smollett	*History and Adventures of an Atom*	(SAA)	Fiction	1768	59,168
30	Smollett	*Humphrey Clinker*	(SHC)	Fiction	1771	150,805
					Sum of word-tokens in the set of Smollett texts:	**1,115,738**
					TOTAL OF WORD-TOKENS IN THE CORPUS:	**5,846,102**

Table 1

Dickens and Smollett stand in contrast in the frequency of superlative forms. In Dickens' 23 texts used in this study, the number of tokens for superlatives amounts to 4,960, whereas Smollett employs them 634 times in his seven works. In the normalised frequency scale per million words, the frequency in Dickens is nearly twice as high as in Smollett: 1,049 versus 568.

Total tokens of superlatives	**14,858**
Tokens in Dickens	11,895
Tokens in Smollett	2,963
Total types of superlatives	**423**
Types in Dickens	415
Types in Smollett	113
Types found *only in Dickens*	310
Types found *only in Smollett*	8
Types found in both Dickens and Smollett	**102**

Rk	Types	Freq.
1	*most RBS*	4825
2	*best*	2093
3	*least*	1779
4	*most*	567
5	*greatest*	487
6	*worst*	374
7	*dearest*	285
8	*nearest*	198
9	*slightest*	196
10	*smallest*	180
.	.	.
.	.	.
423	*wryest*	1

Hapax legomena (164 types): *justest, superlativest, unfortunatest, etc.*

Table 2

With regard to the number of types, 423 different superlative forms are found in total. Among those, a few types are highly frequent such as, *most*, *best*, and *least*, occurring more than one thousand times. Conversely, more than one third of the whole types occur only once. Such hapax legomena include unique words, such as, *superlativest* and *unfortunatest*.

This study deals with a corpus of texts comprising Dickens' and Smollett's major works. Dickens' set includes fifteen "serial fictions", six "sketches", one "miscellany", and one "history". Smollett's contains six "fictions" and one "sketch". The total word-tokens in the corpus amount to 5.8 million, with the Dickens component containing 4.7 million tokens and the Smollett component totalling 1.1 million word-tokens. The present project was initiated as a study based on a comprehensive collection, not a sample corpus, of texts by the targeted authors. Therefore, the imbalance in the number of texts as well as tokens is inevitable. However, due attention will be paid in the choice of variables to minimize a potential effect of the differences in the population of the two sets.

2. Quantitative approaches to style/register variation

Milic (1967) is among the earliest successful specimens of a quantitative description of style. He compared the style of Jonathan Swift with the writings of his contemporaries, with special reference to the relative frequencies of word-classes in the texts and to grammatical features such as seriation and connection. Cluett (1971 & 1976) adopted a similar approach to conduct a diachronic study of prose style across 4 centuries: from 16th to 20th Centuries. Brainerd's works (1979 & 1980) are ambitious attempts to apply discriminant analysis to the question of genre and chronology in Shakespeare plays. Takefuta's (1981) approach to text typology, or register variation, is among the first to successfully employ factor/cluster analysis to the lexical differences between registers. His pioneering work, however, is not widely acknowledged because it was written in Japanese.

Since Burrows (1987) and Biber (1988), it has become popular practice to employ multivariate techniques in quantitative studies of texts. Biber carried out factor analysis (FA) on 67 linguistic features to identify co-occurring linguistic features that account for dimensions of register variation. A series of research based on Biber's Multi- Feature/Multi-Dimensional approach have been successful in elucidating many interesting aspects of linguistic variation, such as diachronic change of prose style, variation within a single author, and differences between conversational styles in British and American English, to give a few examples (Biber & Finegan 1992; Opas-Ha nninen 1996, Watson 1997, Conrad & Bibereds. 2001)

The Biber model is one of the most sophisticated approaches by far. Yet it is not without its critics. Nakamura (1995) raises a major objection. He argues that Biber's variables are "quite arbitrarily selected with no definite criterion and mixed levels" (1995: 77-86). Further, Sigley (1997) notes that almost half of Biber's 67 linguistic features are too rare in texts of 2,000 words.

Burrows (1987), on the other hand, applied a Principal Component Analysis (PCA) to the thirty most common words in the language of Jane Austen. The method demonstrates that differing frequency patterns in these very common words show significant differentiations among Austen's characters, and that the statistical analysis of literary style may lead not only to a deeper understanding of the novel itself but may also contribute to our deeper appreciation of it. In this use of a PCA, the

frequencies of common words are used as variables. The Burrows method seems to have higher replicability and feasibility; since it focuses on common words, most of the variables are frequent enough to produce stable statistical results. In addition, it does not require a multi-layered tagging scheme optimised for Biber's MF/MD approach.

A particular strength of the Burrows methodology is in testing cases of disputed authorship and national differences in the English first-person retrospective narrative, known as 'history'. Among the most successful applications are Burrows (1989, 1992 & 1996), Craig (1999a, b, & c). The Burrows approach or similar methodology has been applied to Bible stylometry. Some scholars like Linmans (1998), Merriam (1998), and Mealand (1999) use Correspondence Analysis (CA) instead of PCA. In the context of text typology, Nakamura (1993) applied CA to the frequency distribution of personal pronouns to visualize association between personal pronouns and 15 text categories in the LOB corpus.

3. Methodology

The present study is different from the Biber and Burrows models in that it extends the range of variables to include low-frequency words, or rare words, by applying CA in the analysis of superlatives. CA is one of the techniques for data-reduction alongside PCA and FA. CA allows examination of the complex interrelationships between row cases (i.e., texts), interrelationships between column variables (i.e., words), and association between the row cases and column variables graphically in a multi-dimensional space. It computes the row coordinates (word scores) and column coordinates (text scores) in a way that permutes the original data matrix so that the correlation between the word variables and text profiles are maximized. In a permuted data matrix, adverbs with a similar pattern of distribution make the closest neighbours, and so do texts of similar profile. When the row/column scores are projected in multi-dimensional charts, relative distance between variable entries indicates affinity, similarity, association, or otherwise between them. One advantage CA has over PCA and FA is that PCA and FA cannot be computed on a rectangular matrix where the number of columns exceeds the number of rows, a concern of the present study. Yet CA can handle such types of a data table with, for example, the row cases consisting of thirty texts and the column variables consisting of hundreds of words.

Rank	Types	Sum	SB	PP	OEP	OT	NN	MHC	OCS	BR	AN	MC	PFI	DS	DC	CHE	BH	HT	LD	RPR	TTC	UT	GE	OMF	ED	SRR	SPP	SFC	SLG	STFI	SAA	SHC
1	most_BBS	4825	357	312	119	110	172	31	125	186	120	222	67	188	207	116	170	68	193	68	57	103	103	154	47	140	529	272	104	163	71	152
2	best	2093	53	101	26	53	135	9	67	111	26	126	21	91	147	71	156	36	124	32	40	57	85	135	32	55	87	34	29	65	17	60
3	least	1779	51	61	20	27	102	15	49	80	41	128	27	113	114	16	112	22	87	32	20	35	41	76	29	57	174	78	30	43	19	66
4	most	567	18	28	18	9	42	4	18	31	17	56	13	26	26	16	35	11	40	12	13	23	19	41	14	12	12	6	2	8	4	13
5	greatest	487	20	14	12	9	17	3	11	14	7	15	3	18	15	14	10	4	12	15	6	20	7	26	20	5	24	53	16	2	17	11	32
6	worst	374	13	12	7	13	29	1	16	34	16	50	1	12	18	26	15	4	15	8	18	21	17	20	5	3	7	3	2	7	1	5	8
7	dearest	285	4	5	4	5	4	0	2	2	0	11	0	53	16	1	29	7	27	1	13	4	3	49	5	3	4	4	0	3	0	6
8	nearest	195	4	10	0	9	19	0	8	17	5	14	4	9	13	1	9	3	14	2	10	10	6	13	2	3	4	3	2	5	1	2
9	slightest	196	24	15	9	8	24	2	12	8	8	8	2	4	5	1	2	4	11	2	1	6	2	14	3	0	0	0	0	1	2	4
10	smallest	180	3	15	6	9	25	0	13	11	12	10	6	4	5	0	4	1	11	4	0	4	1	6	4	5	6	3	1	6	2	3	2
11	biggest	178	7	14	6	4	7	1	2	5	6	10	4	7	8	7	12	3	14	5	2	7	5	14	3	3	17	2	0	3	2	2
12	eldest	172	22	7	0	11	17	4	6	1	0	14	0	7	26	10	8	4	3	2	0	4	1	4	0	5	4	1	2	3	0	1	3
13	youngest	155	20	6	2	7	15	0	1	11	49	0	4	8	6	7	0	0	6	2	1	5	0	10	0	2	1	0	0	0	0	1	4
14	lowest	105	8	1	2	9	0	0	0	3	2	6	4	5	5	0	1	0	8	4	3	6	8	1	10	2	1	2	5	2	5	1	2
15	strongest	84	1	0	0	2	1	0	2	2	5	1	4	3	1	5	3	2	4	2	2	6	2	7	0	3	2	4	3	1	0	2	0
16	deepest	82	4	11	1	2	3	1	3	7	1	11	4	3	6	0	5	1	2	2	0	1	0	2	0	2	0	1	0	1	0	1	0
17	oldest	82	4	5	1	3	4	2	1	3	2	8	0	4	2	4	6	0	4	2	5	8	0	14	0	2	0	2	1	0	0	1	0
18	finest	78	9	3	1	1	8	1	0	7	2	4	2	2	1	0	4	2	4	7	3	3	2	7	1	1	1	1	2	0	0	8	0
19	brightest	66	2	2	1	0	5	1	4	3	4	2	4	4	6	4	11	0	5	0	0	0	0	6	1	0	2	0	0	0	0	0	0
20	earliest	66	2	4	2	0	5	2	4	2	1	4	0	3	3	0	6	1	1	0	0	1	0	10	2	1	1	0	0	0	0	0	0
21	largest	63	1	5	0	0	6	0	2	4	2	1	2	3	2	0	3	1	4	4	1	2	2	3	3	1	1	1	0	0	6	1	1
22	happiest	54	0	5	0	0	4	0	5	3	0	2	0	2	0	0	5	2	2	0	1	0	0	3	0	0	5	3	1	2	1	0	2
23	warmest	53	2	4	1	4	6	0	1	3	1	2	0	3	6	0	2	0	0	1	0	0	1	1	1	0	6	5	0	2	0	0	0
24	strangest	52	3	1	0	0	3	0	1	1	0	1	0	1	6	2	2	1	1	0	0	2	1	4	0	0	0	0	0	0	0	0	1
25	sunniest	50	3	8	1	2	5	2	1	5	2	1	2	2	2	1	6	0	4	0	0	2	0	0	1	0	0	0	0	0	0	1	1
26	uppermost	48	1	3	1	4	6	1	1	2	2	0	1	1	1	0	0	0	3	0	1	1	2	0	0	0	0	1	0	0	1	0	0
27	hottest	46	0	1	0	0	0	1	2	2	3	10	0	2	4	0	2	0	1	2	0	1	0	5	0	0	0	0	0	0	2	0	0
28	commonest	44	2	1	1	1	2	2	1	2	4	4	3	1	3	0	4	1	1	1	0	2	1	2	2	0	0	0	0	0	0	0	1
29	fastest	44	0	4	2	2	5	0	0	5	1	0	0	6	1	0	1	0	4	0	0	1	0	1	0	0	1	0	0	0	0	0	0
30	pleasantest	43	0	1	2	0	2	1	4	3	2	4	0	1	1	0	2	2	2	0	1	0	0	2	2	0	0	1	0	0	0	0	0
31	richest	43	1	1	1	1	4	0	2	2	1	2	0	2	2	0	2	0	2	1	0	0	0	2	1	0	0	0	0	0	0	0	0
32	longest	42	1	3	1	4	2	0	1	6	1	4	0	1	1	0	0	0	4	1	0	0	0	2	0	0	2	1	3	2	1	0	1
33	kindest	41	1	0	0	2	6	0	2	1	0	4	0	0	4	0	1	0	2	1	0	1	0	1	1	0	0	0	0	0	0	0	0
34	wildest	41	1	1	0	2	4	0	0	6	1	2	0	0	4	0	2	0	1	0	0	1	0	4	0	0	1	0	0	0	2	0	0
35	hardest	40	1	0	0	2	6	0	1	3	1	3	0	1	5	1	2	1	1	4	0	1	1	1	2	0	1	0	0	0	1	0	1
36	shortest	40	2	1	0	2	3	0	0	6	1	4	0	1	3	0	1	0	1	1	0	1	0	1	0	0	0	0	0	0	2	0	0
37	sweetest	40	1	1	1	6	0	2	1	1	6	0	9	0	0	1	1	2	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0
38	liveliest	35	1	1	1	1	4	0	0	3	1	3	0	1	3	0	1	0	2	0	0	1	0	1	0	0	0	0	0	0	0	0	0
39	fullest	34	2	3	1	1	4	0	3	3	0	4	1	2	0	0	1	2	4	0	0	1	0	0	1	0	0	0	0	0	0	0	0
40	latest	34	0	1	0	0	3	1	1	2	0	0	0	4	2	0	0	0	1	0	0	1	0	1	0	0	2	0	0	0	0	0	0
41	meanest	33	0	1	0	0	3	1	1	2	0	0	0	2	4	0	0	0	1	0	0	1	0	1	0	0	2	0	0	0	0	0	0
42	wisest	31	0	0	0	0	0	1	0	0	0	2	0	0	4	0	1	0	1	1	0	1	1	2	0	0	2	3	1	0	0	1	0
43	ugliest	30	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0
44	prettiest	29	0	1	0	0	2	0	1	1	0	0	0	2	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
45	sparkliest	28	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
46	farthest	27	1	1	0	2	3	0	0	3	0	1	1	0	0	0	0	0	1	2	0	0	1	1	1	0	0	0	0	0	0	0	0
47	purest	27	1	0	0	2	6	0	1	1	2	0	1	1	1	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	1	1	0
48	darkest	26	2	0	0	2	3	0	0	1	1	2	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
49	fairest	26	0	1	0	0	4	2	1	1	0	2	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	2	2	1
50	truest	26	0	1	0	1	3	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
242	weightiest	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3: Frequency matrix for 242 types of superlatives across 30 texts: raw frequency scores

4. Results

Figures 1 and 2 demonstrate a result of CA based on 242 superlative forms across 30 texts. The solution given as Dimension 1, the most powerful axis, allows quite straightforward interpretation: the horizontal axis of Figure 1 distinguishes between the Dickens and the Smollett sets. It is also interesting that the early Dickensian texts, such as Sketches by Boz, Pickwick Papers, and Nicholas Nickleby, are among the closest to Smollett's texts along the horizontal axis.

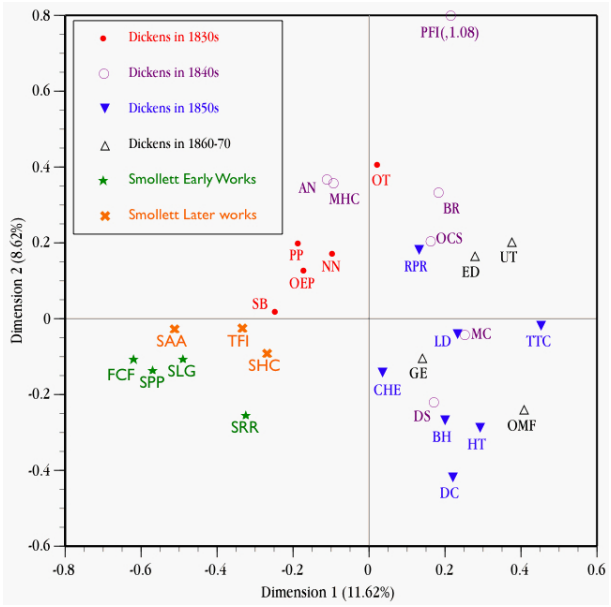


Figure 1: Correspondence Analysis of superlatives in Dickens & Smollett based on 242 types that appear in two or more texts: Text-map showing interrelationships between 30 texts

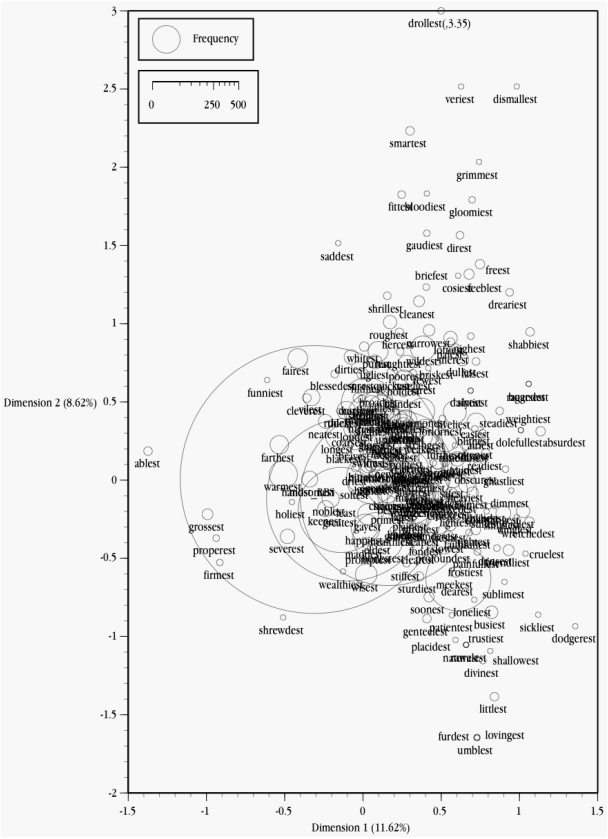


Figure 2: Correspondence Analysis of superlatives in Dickens & Smollett based on 242 types that appear in two or more texts: Word-map showing interrelationships between 242 superlatives

The Dickens corpus is more than four times the size of the Smollett corpus, and the number of types used by Dickens is nearly four times as many as those used by Smollett (see Table 2). It is necessary to ensure that a size factor does not come into play in the outcome of analysis. Figures 3 and 4 are derived from 105 superlatives common to both authors. Despite the decrease in the number of variables from 242 to 105, the configuration of texts and words is remarkably similar to that based on 242 items. Of further interest is that, in each of the two authors' sets, early works tend to have lower scores with later works scoring higher along Dimension 2.

Such result seems to illustrate how the authorial difference and chronology are reflected in the frequency pattern of superlatives in the texts written by Dickens and Smollett. This study might suggest the effectiveness of the stylo-statistical approach based on correspondence analysis of texts.

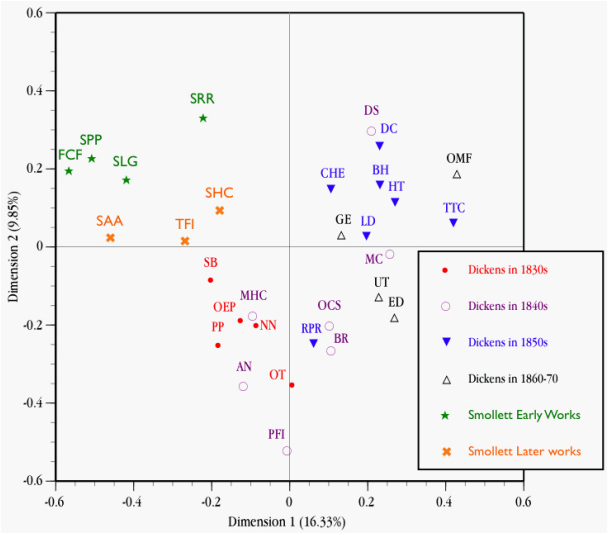


Figure 3: Correspondence Analysis of superlatives in Dickens & Smollett based on the 105 types common to both authors: Text-map showing interrelationships between 30 texts

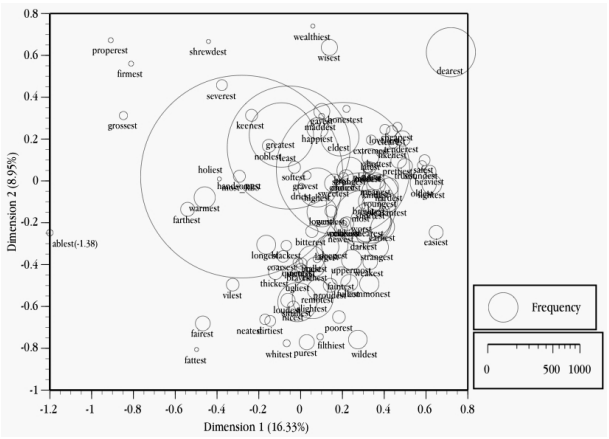


Figure 4: Correspondence Analysis of superlatives in Dickens & Smollett based on the 105 types common to both authors: Word-map showing interrelationships among 105 superlatives

Bibliography

Biber, Douglas. *Variation Across Speech and Writing.* Cambridge: Cambridge University Press, 1988.

Biber, Douglas, and Edward Finegan. "The Linguistic Evolution of Five Written and Speech-Based English Genres from the 17th to the 20th Centuries." *History of Englishes: New Methods and Interpretation in Historical Linguistics.* Ed. Matti Rissanen. Berlin: Mouton de Gruyter, 1992. 668–704.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education Ltd, 1999.

Brainerd, Barron. "The Chronology of Shakespeare's Plays: A Statistical Study." *Computers and the Humanities* 14.4 (1980): 221-230.

Brainerd, Barron. "Pronouns and Genre in Shakespeare's Drama." *Computers and the Humanities* 13.1 (1999): 3-16.

Brook, G. L. *The Language of Dickens.* London: Andre Deutsch, 1970.

Burrows, John F. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press, 1987.

Burrows, John F. ""A Vision" as a Revision?" *Eighteenth-Century Studies* 22.4 (1989).

Burrows, John F. "Computers and the Study of Literature." *Computers and Written Texts.* Ed. Christopher S. Butler. Oxford: Blackwell , 1992. 167-204.

Burrows, John F. "Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative." Ed. Susan Hockey and Nancy Ide. Research in Humanities Computing 4: Selected Papers from the ALLC/ACH Conference, Christ Church, Oxford, April 1992. Oxford: Oxford University Press, 1996. 1-33.

Cluett, Robert. "Style, Precept, Personality: A Test Case (Thomas Sprat, 1635-1713)." *Computers and the Humanities* 5.5 (1971).

Cluett, Robert. *Prose Style and Critical Reading.* New York: Teachers College Press, 1976.

Conrad, Susan, and Douglas Biber, eds. *Variation in English: Multi-Dimensional Studies.* Harlow: Pearson Education Ltd, 2001.

Craig, D. H. "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?" *Literary & Linguistic Computing* 14.1 (1999b): 103-13.

Craig, D. H. "Contrast and Change in the Idiolects of Ben Jonson Characters." *Computers and the Humanities* 33.3 (1999c): 221-40.

Craig, D. H. "Jonsonian Chronology and the Styles of A Tale of a Tub." *Re-Presenting Ben Jonson: Text, History, Performance.* Ed. Martin Butler. London: Macmillan, 1999a. 210-32.

Hori, Masahiro. *Investigating Dickens' Style: A Collocational Analysis.* New York: Palgrave Macmillan, 2004.

Hori, Masahiro. "Collocational Patterns of –ly Manner Adverbs in Dickens." *English Corpus Linguistics in Japan* 15 (2002): 149–163.

Hori, Masahiro . "Collocational Patterns of Intensive Adverbs in Dickens: A Tentative Approach." *English Corpus Studies* 6 (1999): 51–65.

Linmans, A. J. M. "Correspondence Analysis of the Synoptic Gospels." *Literary & Linguistic Computing* 13.1 (1998): 1-13.

Mealand, D. L. "Style, Genre, and Authorship in Acts, the Septuagint, and Hellenistic Historians." *Literary & Linguistic Computing* 14.4 (1999): 479–505.

Merriam, Thomas. "Heterogeneous Authorship in Early Shakespeare and the Problem of Henry V." *Literary & Linguistic Computing* 13.1 (1998): 15–28.

Milic, Louis Tonko. *A Quantitative Approach to the Style of Jonathan Swift.* The Hague: Mouton, 1967.

Nakamura, Junsaku. "Text Typology and Corpus: A Critical Review of Biber's Methodology." *English Corpus Studies* 2 (1995): 75-90.

Nakamura, Junsaku. "Statistical Methods and Large Corpora: A New Tool for Describing Text Types." *Text and Technology: In Honour of John Sinclair.* Ed. Mona Baker, Gill Francis and Elena Tognini-Bonelli . Amsterdam: John Benjamins, 1993. 293–312.

Opas-Ha nninen, Lisa Lena. "A Multi-Dimensional Analysis of Style in Samuel Beckett's Prose Works." *Research in Humanities Computing 4: Selected Papers from the ALLC/ACH Conference, Christ Church, Oxford, April 1992* . Ed. Susan Hockey and Nacny Ide. Oxford: Oxford University Press, 1996. 81-114.

Sigley, Robert. "Text Categories and Where You Can Stick Them: A Crude Formality Index." *International Journal of Corpus Linguistics* 2.2 (1997): 199–237.

Sørensen, Knud. *Charles Dickens: Linguistic Innovator.* Aarhus: Aarhus Universitet, 1985.

Takefuta, Y. *Kompyuta no mita gendai eigo: bokyaburari no kagaku [The Computer Analysis of the Contemporary English Language: A Quantitative Study of Vocabulary].* Tokyo: Educa, 1981.

Watson, Greg. *Doin' Mudrooroo: Elements of Style and Involvement in the Early Prose Fiction of Mudrooroo.* Publications in the Humanities, No 19 . Joensuu, FI:: University of Joensuu, 1997.

Researching e-Science Analysis of Census Holdings: The ReACH project

Melissa Terras *(m.terras@ucl.ac.uk)*
School of Library, Archive and Information Studies
University College London

e-Science technologies have the potential to enable large-scale datasets to be searched analysed, and shared quickly, efficiently, and in complex and novel ways. So far, little application has been made of the processing power of grid technologies to humanities data, due to lack of available large scale datasets which would warrant such high performance computing, and little understanding of or access to e-Science technologies. The ReACH workshop series, funded by the UK's Arts and Humanities Research Council, was established in June 2006 at University College London to investigate the potential application of e-Science and high performance computing technologies to a large dataset of interest to historians, humanists, digital consumers, and the general public: historical census records.

The ReACH series consisted of various workshops undertaken over the summer of 2006 to investigate the academic, technical, and managerial aspects that would have to be taken into account in order to set up a large scale project which would utilise UCL's high performance computing facilities to analyse large scale historical census datasets from the UK's National Archives, in conjunction with the genealogy firm, Ancestry. By undertaking a scoping study in this manner, it was hoped to determine the academic merits of such a proposal: it may be feasible to undertake this analysis, but would it be useful to historical researchers? What would the analysis do? What would the technical implementation of such a project involve? What staffing and funding costs would be required? The workshop series featured input from various project partners, and interdisciplinary experts, to ascertain whether a full scale project would be worthwhile to undertake. Moreover, the workshop series aimed to ascertain if and how e-Science (defined as "a specific set of advanced technologies for Internet resource-sharing and collaboration: so-called grid technologies, and technologies integrated with them, for instance for authentication, data-mining and visualization. (AHRC ICT 2006)") can be applied to the arts and humanities.

Public interest in historical census data is phenomenal, as the overwhelming response to mounting the 1901 census online at The National Archives demonstrates (Inman, 2002). Yet the data is also much used for research by historians (see Higgs 2005 for an introduction). There are many versions of historical census datasets available, covering a variety of aspect of the census, and digitised census records are one of the largest digital datasets available in arts and humanities research. In the Arts and Humanities Data Service repository collection alone there are currently 155 datasets pertaining to historical census data (from the UK and abroad) created for research purposes (AHDS 2006). Commercial firms dealing (or having dealt) in genealogy information (such as Ancestry[1], Genes Re-united[2], QinetiQ [3], British Origins[4], The Genealogist[5], and 1837Online[6]) have digitised vast swathes of historical census material (although to varying degrees of completeness and accuracy). There is much interest from the historical community in using this emerging data for research, and developing tools and computational architectures which can aid historians in analysing this complex data (see Crocket, Jones and Schürer (2006) for an advanced proposal regarding the creation of a longitudinal database of English individuals and households from 1851 to 1901, see also the work of the North Atlantic Population Project[7]). However, there have been few opportunities for the application of high performance computing to utilise large scale processing power in the analysis of historical census material, especially analysing data across the spectrum of census years available in the UK (7 different censuses taken at 10 year intervals from 1841-1901). Although certain digitized datasets of the UK census are in the public domain (1881[8]) most were digitized by commercial companies and are unavailable to the academic researcher. Most historians do not have access to, or do not know how to use, high performance computing facilities.

The aim of the ReACH series was to bring together disparate expertise in Computing Science, Archives, Genealogy, History, and Humanities Computing, to discuss how e-science scale techniques could be applied to be of use in the historical research community. The project partners each brought various expertise and input to the project:

- UCL School of Library, Archives and Information Studies[9], who have expertise in digital humanities and advanced computational techniques, as well as digital records management,
- The National Archives[10], who select, preserve and provide access to, and advice on, historical records, e.g. the censuses of England and Wales 1841-1901 (and also the Isle of Man, Channel Islands and Royal Navy censuses)
- Ancestry.co.uk[11], who own a massive dataset of census holdings worldwide, and who have digitized the censuses of England and Wales under license from The National Archives. The input of Ancestry was central to this research to gain access to the complete range of UK census years in digital format.

- UCL Research Computing[12], the UK's Centre for Excellence in networked computing, who have extensive high performance computing facilities available for use in research.

The project aimed to investigate the reuse of pre-digitised census data: presuming there was not funding available to be in the business of digitisation of other record data for any pilot project. The project also wished to investigate the use of commercial datasets (as many of the large census data sets are owned by commercial firms: in this case, Ancestry), and the licensing and managerial issues this would raise for future projects. The project also wanted to establish how feasible, and indeed useful, undertaking such an analysis of historical census data would be.

The results of the well attended workshop series was a sketch for a potential project, and recommendations regarding the implementation of e-science (high performance computing) technologies in this area. However, at this time, it was not thought possible to pursue the potential project at this time in the following e-Science call which emanated from the AHRC in October 2006 due to a variety of reasons which are elucidated in this paper. Reasons for not taking the project forward at this time were not technical or managerial, but historical: it will be a few years before all the digitized data required to make this project a success will be available (or be of high enough quality, see Holmes 2006). Nevertheless, the scoping nature of this project did highlight interesting aspects of the application of high performance computing to humanities data: discussing the nature, size and quality of humanities datasets (as opposed to scientific datasets), and managerial and technical expertise in data management, security, and licensing. Importantly, the nature of working with a commercial company on their sensitive data was also explored from a legal aspect, highlighting issues regarding use and reuse of digital data for the arts and humanities: who “owns” resulting datasets from collaborative projects?

This paper describes the methodology of the workshops, reporting on suggestions made during the series regarding potential applications of high performance computing which would benefit academic historians, sketching out a future project regarding how historical census material can be analysed utilising high performance computing, and extrapolates recommendations that can be applied in general to the use of e-Science and high performance computing in the arts and humanities research sectors.

---

1. <http://www.ancestry.com/>
2. <http://www.genesreunited.co.uk/>
3. <http://www.qinetiq.com/>
4. < http://www.origins.net/BOWelcome.aspx>
5. <http://www.thegenealogist.co.uk/>
6. <http://www.1837online.com/>
7. <http://www.nappdata.org/napp/>
8. The 1881 Census for England and Wales, the Channel Islands and the Isle of Man (Enhanced Version) was deposited in the Arts and Humanities Data Service repository by K. Schürer (University of Essex. Department of History) in 2000, and is available from <http://www.ahds.ac.uk/catalogue/collection.htm?uri=hist-4177-1>
9. <http://www.slais.ucl.ac.uk/>
10. <http://www.nationalarchives.gov.uk/>
11. <http://www.ancestry.co.uk/>
12. <http://www.ucl.ac.uk/research-computing/>

Bibliography

Arts and Humanities Data Service (AHDS). Cross Search Catalogue. 2006. Accessed 2006-10-31. <http://www.ahds.ac.uk/catalogue/search.htm?nq=n&q=census&s=all&coll=y&item=y>

Arts and Humanities Research Council (AHRC). AHRC ICT Programme Activities and Services. 2006. Accessed 2006-11-13. <http://www.ahrcict.rdg.ac.uk/activities/e-science/background.htm>

Crocket, A., C. E. Jones, and K. Schürer. The Victorian Panel Study. Report Submitted to the ESRC (Award Ref: RES-500-25-5001), May 2006. 2006.

Higgs, Edward. *Making Sense of the Census Revisited: Census Records for England and Wales 1801-1901: A Handbook for Historical Researchers.* London: Institute of Historical Research, 2005.

Holmes, R. "The Accuracy and Consistency of the Census Returns for England 1841-1901 and their Indexes." M.A. Dissertation. School of Library, Archive and Information Studies, University College London, 2006.

Inman, Phillip. "Genealogy." *The Guardian* (Thursday September 26, 2002). Accessed 2006-11-03. <http://www.guardian.co.uk/internetnews/story/0,,798781,00.html>

ADHO Panel: Beyond Text

John Unsworth *(unsworth@uiuc.edu)*
University of Illinois at Urbana-Champaign
Kevin Franklin *(kfranklin@hri.uci.edu)*
University of California Humanities Research Institute (UCHRI)
Matt Kirschenbaum *(mkirschenbaum@gmail.com)*
University of Maryland
Lev Manovich *(manovich.lev@gmail.com)*
University of California, San Diego
Catherine Plaisant *(plaisant@cs.umd.edu)*
University of Maryland
University of California, San Diego

This panel will engage in a discussion of work being done in the humanities that is computational but not literary/linguistic, or at least not primarily focused on text. The panel will also discuss the possibilities for collaboration across the different paradigms of humanities computing/new media/visual culture/grid computing/critical studies.

Moderator: John Unsworth

Participants:

Kevin Franklin

Kevin Franklin is executive director of the University of California Humanities Research Institute (UCHRI) and a former deputy director of the San Diego Supercomputer Center. He serves as co-chair for the Humanities, Arts and Social Science Research Group for the Global Grid Forum and is a member of the UC Humanities, Arts and Social Science Technology Council and the Worldwide University Network Grid Advisory Committee. Franklin coordinates UCHRI research and development activities at the interface of the humanities, arts, science and technology.

Lev Manovich

Lev Manovich (<http://www.manovich.net>) is a Professor in the Visual Arts Department, University of California, San Diego where he teaches courses in new media art and theory. He is the author of The Language of New Media (The MIT Press, 2001), Tekstura: Russian Essays on Visual Culture (Chicago University Press, 1993) as well as many articles which have been published in 28 countries. Manovich is in demand to lecture on new media; since 1999 he delivered over 180 lectures in North and South America, Europe, and Asia. His awards include Mellon Fellowship and Guggenheim Fellowship (2002-2003).

Matt Kirschenbaum

Matthew G. Kirschenbaum is an Associate Professor in the Department of English at the University of Maryland and Associate Director of the Maryland Institute for Technology in the Humanites (MITH), an applied thinktank for the digital humanities. He is also an affiliated faculty member with the Human-Computer Interaction Lab at Maryland, and a Vice President of the Electronic Literature Organization. Kirschenbaum specializes in digital humanities, electronic literature and creative new media (including games), textual studies, and postmodern/experimental literature. He has a Ph.D. in English from the University of Virginia, and was trained in humanities computing at Virginia's Electronic Text Center and Institute for Advanced Technology in the Humanities (where he was the Project Manager of the William Blake Archive). His dissertation was the first electronic dissertation in the English department at Virginia and one of the very first in the nation.

Catherine Plaisant

Dr. Catherine Plaisant is Associate Research Scientist at the Human-Computer Interaction Laboratory of the University of Maryland Institute for Advanced Computer Studies. She earned a Doctorat d'Ingenieur degree in France in 1982. In 1987 she joined Professor Ben Shneiderman at the Human-Computer Interaction Laboratory. She enjoys most working with multidisciplinary teams on designing and evaluating new interface technologies that are useable and useful. Her research contributions range from focused user interaction techniques (e.g. Excentric Labeling) to innovative visualizations (such as LifeLines for personal records or SpaceTree for hierarchical data exploration) and interactive search interface techniques such as Query Previews. Those interaction techniques have been carefully validated with user studies and are finding applications in industry and government information systems and digital libraries. She has written over 90 refereed technical publications on the subjects of information visualization, digital libraries, universal access, image browsing, input devices, online help, home automation, network management, telemedicine etc. She recently co-authored with Ben Shneiderman the 4th Edition of Designing the User Interface, one of the major books on the topic of Human-Computer Interaction.

Second Life for Museums and Archeological Modeling

Richard Urban *(rjurban@uiuc.edu)*
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Michael Twidale *(twidale@uiuc.edu)*
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Paul F. Marty *(marty@ci.fsu.edu)*
School of Information
Florida State University

Museums have been exploring the use of multi-user virtual environments (MUVE) for more than a decade, often in the form of proprietary virtual worlds built for select audiences such as teachers and students. Since launching in 2003, the online virtual community of Second Life (<http://www.secondlife.com>) has attracted over one million dedicated 'residents' who are laying the foundations for widespread adoption of MUVE. In many ways, the growth of MUVE mirrors the growth of the Web. New technologies transition from small scale prototypes constructed by researchers at great expense to large scale, rapidly growing mainstream products and services available to the general public. These products are not only used by many people, but are co-created by them. With the Web this was a matter of using hypertext to create websites, initially inspired by various genres of print media, and soon evolving their own genres. In the case of MUVE like Second Life (SL), residents can create 3D artifacts, buildings, and social spaces where people interact. The social nature of Second Life is a critical component of understanding how it is, can, and should be used.

Already we can see a wide range of museum-like activities occurring in Second Life. A recent survey identified over 150 museums, galleries or museum-like activity spaces in Second Life. These museums offered visitors opportunities to view collections of real-life and fictional spacecraft, digitized versions of real-life artworks, exhibitions of artworks created in Second Life, living history reenactments, or archaeological monuments (Urban, Marty & Twidale, 2007). Reflecting the patterns of development of the early World Wide Web, most of these SL museum spaces have been created by enthusiasts - residents who are not affiliated with real-life museums.

Game studies researchers have suggested that the behaviors of players in other massively multi-player online games (MMOG) such as *World of Warcraft* are blurring the boundaries between work, play and learning (Castronova, 2006; Steinkuehler & Williams, 2006; Yee, 2006). Like other passionate user communities, Second Life residents are turning the virtual world into a 'third place' where they can engage in serious leisure pursuits. Museums and cultural heritage institutions are not unfamiliar with serious leisure as they often engage with amateur archaeologists, natural scientists, living-history re-enactors or family historians. Unlike traditional museum audiences, these individuals are "involved participants rather than consumers" (Orr, 2006; Stebbins, 1992).

The degree to which individual residents decide to place themselves on a continuum from pure work to pure play can affect the results of their activities in Second Life. Some residents seek to evoke a particular historic place, while others engage in the serious work of conducting research, visiting the real-life places and artifacts upon which their Second Life representations are based. Some residents see Second Life as primarily a social space that requires a lower degree of authenticity and accuracy. As long as the stage can provide the necessary background, it can facilitate social interactions (DiBlas, 2005). Other residents are using Second Life as a new expressive medium that allows them to create new artworks or to accurately represent cultural artifacts as best they can.

As residents create museum-like activities, humanities scholars are exploring Second Life as a virtual classroom. The New Media Consortium (<http://www.nmc.org/sl/>) has constructed a virtual campus where classes are held in "outdoor" amphitheaters, faculty and students are staging plays, poetry readings and providing space for digital artists to create new works. The Stanford Humanities Lab (<http://shl.stanford.edu/>), the Humanities, Arts, Science and Technology Advanced Collaboratory (HASTAC - <http://www.hastac.org/>) have also staked some ground in Second Life.

Observations of how Second Life residents are currently engaging in serious leisure activities around museums and archaeological models can inform many kinds of research. We can study the early stages of a new online social medium being co-created and evolving into a new form, just as the early days of the web led to the development of new forms of creating, sharing and manipulating information, including cultural materials. We can also use the work of these technological pioneers to inform ways to create new learning spaces for students. Second Life also affords researchers and students an opportunity to demonstrate what is possible when more rigorous methods are applied. While many projects have created models

using highly accurate rendering systems, Second Life can add value to this work by providing tools for social engagement for a broader audience (Di Blas, 2003, 2005; Eiteljorg, 2004). The serious leisure activities of Second Life residents also suggest a more open and inquiry based approach to learning. Instead of presenting students with already completed models, Second Life can engage students through co-creation and collaborative discovery.

This poster will present examples of museum-like spaces and activities taking place in Second Life with a particular focus on archeological-themed representations. It will illustrate how Second Life museums are largely the product of resident's serious leisure activities. Using these activities as an example of what is possible it will explore how early attempts at teaching in Second Life might be informed by resident's serious leisure activities.

Figure 1: Themiskyra Throne Room

Bibliography

Castronova, Edward. *Synthetic Worlds: The Business and Culture of Online Games.* Chicago: University of Chicago Press, 2006.

Di Blas, Nicoletta, Evelyne Gobbo, and Paolo Paolini. "The SEE Experience: Edutainment in 3D Virtual Worlds." *Museums and the Web 2003: Selected Papers from an International Conference.* Ed. David Bearman and Jennifer Trant. Pittsburgh, PA: Archives & Museum Informatics, 2003. 173-182. <http://www.archimuse.com/mw2003/papers/diblas/diblas.html>

Di Blas, Nicoletta, Evelyne Gobbo, and Paolo Paolini. "3D Worlds and Cultural Heritage Realism vs. Virtual Presence." *Museums and the Web 2005: Proceedings.* Ed. David Bearman and Jennifer Trant. Toronto, Ontario: Archives & Museum Informatics, 2005. <http://www.archimuse.com/mw2005/papers/diBlas.html>

Eiteljorg II, Harrison. "Computing for Archeologists." *A Companion to Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Malden, MA: Blackwell Publishing Ltd, 2004. 20-30.

Orr, Noreen. "Museum Volunteering: Heritage as 'Serious Leisure' ." *International Journal of Heritage Studies* 12.2 (2006): 194-210.

Stebbins, Robert A. *Amateurs, Professionals and Serious Leisure.* Montreal: Queens University Press, 1992.

Steinkueler, Constance, and Dmitri Williams. "Where Everybody Knows Your (Screen) Name: Online Games as 'Third Places' ." *Journal of Computer-Mediated Communication* 11.4 Article 1 (2006).

Urban, Richard, Paul F. Marty, and Michael B. Twidale. "A Second Life For Your Museum: 3D Multi-user Virtual Environments and Museums." *Museums & the Web 2007* (Forthcoming).

Yee, Nick. "The Labor of Fun: How Video Games Blur the Boundaries of Work and Play ." *Games and Culture* 1 (2006): 68-71. <http://www.nickyee.com/pubs/Yee%20-%20Labor%20of%20Fun%202006.pdf>

Re-imag[en]ing Cervantes' Don Quixote: a Multi-layered Approach to Editing Visual Materials in a Hypertextual Archive

Eduardo Urbina *(e-urbina@tamu.edu)*
Cervantes Project &
Center for the Study of Digital Libraries
Texas A&M University

Fernando González Moreno
Cervantes Project &
Center for the Study of Digital Libraries
Texas A&M University

Richard Furuta *(furuta@cs.tamu.edu)*
Cervantes Project &
Center for the Study of Digital Libraries
Texas A&M University

Steven E. Smith
Cervantes Project &
Center for the Study of Digital Libraries
Texas A&M University

Jie Deng
Cervantes Project &
Center for the Study of Digital Libraries
Texas A&M University

Stephanie Elmquist
Cervantes Project &
Center for the Study of Digital Libraries
Texas A&M University

Sarah Tonner
Cervantes Project &
Center for the Study of Digital Libraries
Texas A&M University

It is often stated by critics that the Quixote is a theatrical, graphic, and visual book. Thus, visual representations, like theatrical performances, popular iconography and book illustrations, have been recognized as significant contributions to the understanding of Cervantes' masterpiece.[1] [2] Nevertheless, the thousands of woodcuts, engravings, etchings, drawings, and lithographs that have accompanied the text are, for the most part, a little known interpretative tradition, and a much neglected critical and artistic treasure.[3]

Obstacles, such as the difficulty of accessing rare books, have prevented the illustrative tradition from being appreciated by scholars, students, and users in general. In 2001 the Cervantes Project (CP) started the creation of a hypertextual archive to include digital images of the illustrations taken from over 500 of the most significant editions to form the textual iconography of the Quixote (as permitted by copyright limitations). Our main objectives are to make the illustrations more accessible and to establish their contribution to the reception and interpretation of the text. At the present time, the archive has digitized, and made available online more than 7,000 images, supported by a fully searchable database and complemented by rich metadata and innovative visualization tools.[4] (Figure 1 shows the reader's Web-based access to the collection.)

The multidisciplinary approach of our project enables scholars to go beyond the literary aspect of Cervantes' works. As an invaluable pictorial depository, we emphasize supplying information regarding the historic value and artistic significance of the images. The hermeneutic and aesthetic values of each individual image have been carefully examined by art historians and the results incorporated in the archive as scholarly commentary. Additionally, we include biographical commentary about artists and engravers. These rich scholarly commentaries will help to boost the study of book illustration art, which has been to date secondary in Art History, in aspects such as the evolution of techniques, from the first woodcuts (early 17th century) to modern mechanical offset (20th century), and the influence or achievement of an individual engraver, illustrator, or lithographer. We associate the text and the images through a taxonomy of episodes, adventures and narrative motifs for both parts of the Quixote. The 308 taxonomic elements in which we have divided the 126 chapters of the text encompass as descriptive categories the totality of the tradition of illustrations present in the editions in our collection. (See Figure 2 for an example selection from the taxonomy.)

The reader's interface offers access to all the levels of editorial annotation about the artists, textual location, image information and technique, aesthetic and textual commentary. In addition, users can obtain, through links, specific information related to the narrative, biographical information about the creators, and the meaning of engraving techniques. We are in the process of incorporating these and other categories into our current search tool, as well as developing a controlled vocabulary about themes and characters in the Quixote and about generic area content

(flora, fauna, architecture, music, etc.) to facilitate the use and exploration of the images in the archive by researchers other than literary scholars. (Figure 3 shows the advanced search interface.)

As a multidisciplinary project that involves library staff digitizing the images, computer scientists coding the applications, Hispanic scholars providing textual editing and edition metadata, art historians examining the image metadata, and finally the general public traversing the archive, it is critical to have an efficient production line and a strategy for consistently sharing information among the different parties. After several attempts, we reached a work flow that we found to work efficiently. The procedure starts with Hispanic scholars providing bibliographic information and specifying the editions to be digitized, the pages where the images reside, and the naming scheme for the electronic files. This information is recorded on a work sheet and sent to library personnel. Upon receiving such a work sheet, the library staff responsible for the digitization reviews the information, clarifies any possible discrepancies, digitizes the illustrations, and saves them on a library server as archival TIFF files. A copy of the images is then transferred to the development site where a computer scientist processes them into JPEGs, whose sizes are more appropriate for Web traffic. Afterwards, the references to the JPEGs are added to the database. At that point, the interfaces automatically update and these images become available online. The art historian in charge of image metadata has access to the original illustrations and then remotely enters the information using the online Web form. As soon as these commentaries are finalized, they are published online alongside the images.

A Web-based interface is provided for maintenance of the collection. The collection maintenance application supports management functions such as inserting, deleting, or modifying, project-wide data. The editor's Web form also offers a typology of illustrations divided into 17 categories, i.e., frontispiece, vignette, portrait, map, etc., and a 17 item index of engraving techniques. Much of the design effort in preparing this interface has focused on specification of the most appropriate handling of default values and modes. Using an iterative development methodology, we have focused on the art historian's tasks, refining the means for inheriting and modifying values between entries with the goal of focusing his data-entry activities on differences rather than similarities. The taxonomic and typological categories are part of the rich image metadata approach we have developed to capture and make accessible the variety of image content present in the iconography of the Quixote. They are of value both to collection editor and reader as they are part of the editor's metadata template and of the reader's advanced search tool. (Figure 4 shows the tools used by the editor in specifying metadata.)

The readers' Web-based interface, introduced earlier, which is separate from the collection maintainer's interface, also has undergone refinement during the project's development. At present, it provides a four-layered browsing design: a) an edition index; b) a thumbnail overview of one entire edition; c) a low-resolution image together with meta-data; and d) a high-resolution image. One interesting characteristic in the public interface is the provision of uncommon relationships among the artifacts. One particular example is the ancestor-descent relationship. For instance, an engraver is often related to another as his master or disciple, or even father or son; one illustration or a set of illustrations is inherited from an earlier edition as simply reprints or sometimes after re-engraving. In cases where ancestral and/or familial ties exist, we provide a description of the ties, as well as physical links pointing to the referred artifacts for the purpose of more in-depth investigation. Navigation along the ties provides an intuitive way to look closely at the propagation of engraving skills from one person to another, the effect of the technology evolution upon art works as reflected in re-engravings, and the significance, elegance, and popularity of certain artists or illustrations serving as the archetypes for imitation. At present, the links are manually encoded on a one-by-one basis. We are in the process of setting up a keyword-anchor list for automatic linking.[5] (Figure 5 shows the reader's views of the collection and metadata.)

The project's metadata is imbedded into a Dublin Core framework. We use a MySQL server as the database core to store the metadata and the references to image files. The readers' and maintainers' Web-based applications occupy a display tier on top of the database tier.

The availability of the archive will contribute to a more complete understanding and appreciation of Cervantes' novel by initiating new explorations from many perspectives: textual, artistic, critical, bibliographical, and historical. In particular, we provide resources and assistance to examine the reception and evolution of the Quixote's readings across time, culture, audience, and milieu. Furthermore, the images can be grouped according to several layers of content to respond to the users' need for information selection of a specific critical focus, i.e, art, geography, history, etc. This is achieved by 1) cataloging each image using a comprehensive taxonomy of the episodes and adventures, 2) including multiple levels of textual annotation about each individual illustration, including technical, historical, and artistic information, and 3) providing descriptive and critical commentary related to the content and textual context of the illustrations. The result is a multi-layered and multi-directional collection of digital objects recombining bibliographical, textual and visual materials (edition-text-image-metadata), and the development of a rich hypertextual archive encompassing a new form of critical

apparatus in which the illustrations are newly re-contextualized and re-imag[in]ed.

Figure 1: The front page of the collection (left) and the interface to the iconography collection (right).

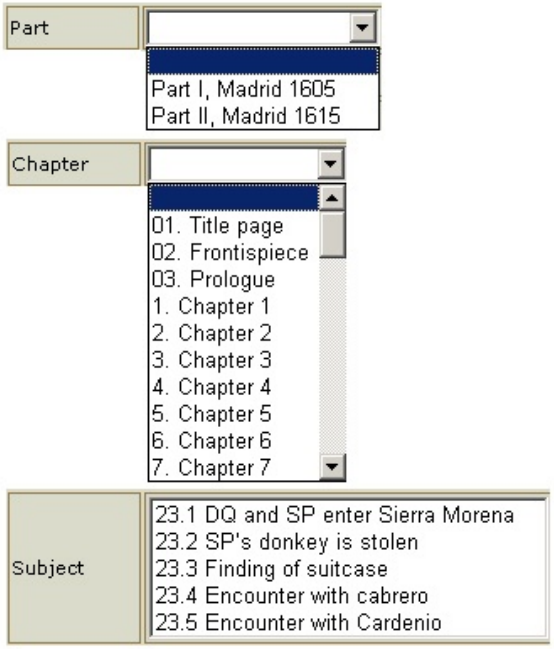


Figure 2: An example set of menus showing the selection of an entry in the thematic taxonomy.

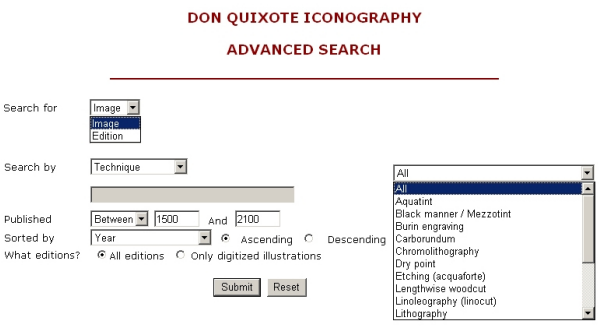


Figure 3: The reader's advanced search interface.

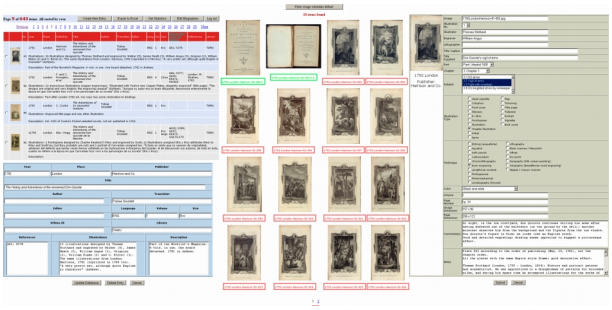


Figure 4: Screen shots from the editor's metadata specification toolkit. At top left is the editor's view of the collection, which resembles the reader's. In addition to allowing selection of editions, this screen allows editing of edition metadata, lower left. Selecting an edition brings up the middle display, which flags images still requiring metadata entry. The metadata is then provided using the screen on the right.

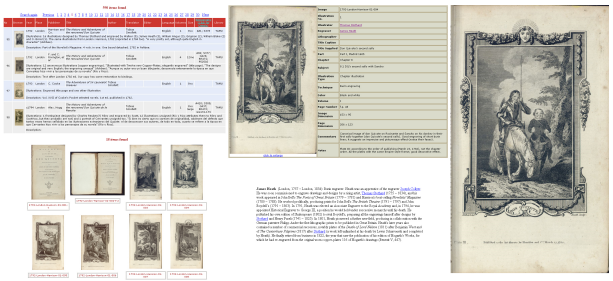


Figure 5: The reader's view of image magnifications and metadata. At top left is the collection view, seen before. Selecting an edition from the collection brings up its images, lower left. Next, selecting an image brings up more detail about an image, including its metadata, center top. Here, the reader can request detail about persons mentioned in the metadata (center bottom) and an additional image magnification (right).

---

1. Funding and support provided by the National Endowment for the Humanities, the Cátedra Cervantes at the University of Castilla-La Mancha, Grupo Santander, Texas A&M University, and the Cushing Memorial Library, Texas A&M University Libraries.

2. See for instance, Jean Canavaggio, *Don Quichotte du livre au mythe* (Paris: Librairie Arthème Fayard, 2005) and José Manuel Lucía Megías, "Don Quijote de la Mancha: del libro al mito," *En torno al Quijote: Adaptaciones, imitaciones, imágenes y música* (Madrid: Biblioteca Histórica Municipal de Madrid, 2005) 27-55.

3. Much has been done in conjunction with the 400th centenary celebration to remedy this situation; in addition to our own project and digital archive at `<http://cervantes.tamu.edu/>` (accessed November 12, 2006) see the Banco de imágenes del *Quijote* directed by J.M. Lucía Megías at `<http://www.qbi2005.com/>` (accessed November 12, 2006).

4. Fernando González Moreno et al, "La colección de *Quijotes* ilustrados del Proyecto Cervantes: Catálogo de ediciones y archivo digital de imágenes," *Cervantes: Bulletin of the Cervantes Society of America* 25.1 (2005) [2006]: 79-104 and E. Urbina et al, "Visual Knowledge: Textual Iconography of the Quixote, a Hypertextual Archive." " Literary & Linguistic Computing" 21.2 (2006): 247-58.

5. Neal Audenaert et al, "A General Framework for Feature Identification," *Digital Humanities 2006 Conference Abstracts*, Association for Digital Humanities Organization International Conference, Université Paris-Sorbonne, (2006) 5-9, and E. Urbina et al, "Textual Iconography of the Quixote: a Data Model for Extending the Single-Faceted Pictorial Space into a Poly-Faceted Semantic Web," *Digital Humanities 2006 Conference Abstracts* (2006) 215-20.

Automatic Techniques for Generating and Correcting Cultural Heritage Collection Metadata

Antal van den Bosch *(Antal.vdnBosch@uvt.nl)*
Tilburg University
The Netherlands
Caroline Sporleder *(C.Sporleder@uvt.nl)*
Tilburg University
The Netherlands
Marieke van Erp *(M.G.J.vanErp@uvt.nl)*
Tilburg University
The Netherlands
Stephen Hunt *(S.J.Hunt@uvt.nl)*
Tilburg University
The Netherlands

In cultural heritage domains the central element of interest in terms of information and knowledge, but also in terms of objective and subjective value, is the *object*. Text is an important medium in retrieving objects, or information about these objects. Even in cases of non-textual cultural heritage collections, text plays a crucial role, since text is the pervasive medium used for metadata, next to numeric identification codes and measurements (Chapman, 2005). At the same time, language is a noisy and ambiguous medium, and usage of textual metadata ultimately implies robustness to noise and proper disambiguation. If the goal is to automate the process of searching and retrieving information about objects from digitized metadata, *language technology* can aid in disambiguating, translating, and correcting textual metadata (Bontcheva et al., 2002). In this paper we describe ongoing work in the MITCH (Mining Information in Texts from the Cultural Heritage) project in which we employ diverse language technologies to enrich textual metadata of a large object collection in the natural history domain.

Textual metadata can take many forms. Many document collections are made searchable through the use of index systems (pre-digital as well as digital) that use keywords, titles, and proper names (e.g. of authors). Alternatively, objects may be described more verbosely in annotations, captions,

explanatory texts for educational purposes or exhibitions, or scholarly publications. A third type of textual metadata is the thesaurus or ontology, which aims to capture a standardized list or set of terms that occur in the domain, and possibly some basic relations between them such as synonymy, hyponymy, and hyperonymy, or more domain-specific relations. Fourth, auxiliary resources peripheral to the domain may exist that contain information of overlapping interest, such as lists of geographical names.

There is a need, clearly expressed by many cultural heritage institutions, for digitizing the metadata of cultural heritage collections, in order to make them searchable and accessible at unprecedented scales. After the basic metadata is digitized, allowing simple keyword-based search, an important next step is the *enrichment*phase, the result of which enables more complex and advanced types of search that better resemble the typical questions of scholarly researchers. The key step in the enrichment phase is to link all resources into a single network. By establishing all meaningful links, a particular object effectively becomes a node in this network, meaningfully connected to all other entities that have a direct relation to that object, such as the artist who created it, or the location in which it was found.

By traversing the network, a search engine can explore complex relations among sets of objects, and unions and intersections of such sets, allowing, for instance, biologists to perform large-scale searches for patterns and trends, such as the occurrence of animals in a certain family, in particular stretches of time (e.g. the last century) and space (e.g. the Amazon rain forest). To achieve this goal, we develop language technology of two types: (1) preprocessing technology for the automatical correction and normalization of our textual resources, and (2) metadata enrichment technology for identifying relevant concepts in all textual metadata, and linking all resources into a connected, searchable network.

Preprocessing: Automatic metadata cleanup

The first step in enriching textual metadata is to make sure that the textual metadata is as correct and consistent as possible. In Sporleder et al. (2006a, 2006b) we describe two complementary methods. The first method, *horizontal correction*, aims at correcting inconsistent values in a database record based on the other values in that record, and all other records in the same database, while the second method, *vertical correction*, focuses on values which were entered in the wrong column. Both methods are language-independent, rely solely on the database itself, and are fully automatic. Their novelty lies in the fact that they can deal with textual material in database cells, where most traditional error correction methods only work for numeric (Hawkins, 1980) or atomic categorial data (Knorr and Ng, 1998).

In a range of experiments on a natural history specimen database we observed that the two methods can effectively zoom in on the errors in the database. Instead of having to manually inspect all 16 thousand records in our database, each characterized by 47 attributes (columns), the methods only flag in the order of several hundreds of errors. The errors flagged tend to capture well above 90% of all actual errors, and typically raise about as many false alarms as proper hits. The method is implemented into a prototype system that draws the attention of the expert user, with simple visual means, to alleged errors in the database. The user is able to approve and effectuate the correction, or to enter another correction, or to dismiss the error message.

Other cleanup methods currently under investigation are the identification and normalization of different language versions of the same location names (Van Erp, 2006), the disambiguation of proper names referring to different entities (Bagga and Baldwin, 1998) and the normalization of time expressions.

Automatic metadata generation

After resources have been cleaned and normalized, similar methods can be used to generate more metadata, in order to improve access to the core data. Again, we follow a strategy which is fully automatic and completely self-sufficient, not relying on any source other than the data itself, because in the typical cultural heritage domain there is little useful background resource besides the data itself.

As a first example, Sporleder et al. (2006c) describe a *bootstrapping* method to generate lists of proper names of persons, locations, and taxonomic animal names from a specimen database. In general, an object database offers its own structured annotation by means of its column titles, such as "location" (of the place where an animal specimen was found). Gathering the contents of cells in a particular column results in a list of named entities of that particular type, that may be found elsewhere in the same database, e.g. in comment fields, or in other textual metadata resources such as field books or scholarly articles. These texts can then be linked to the entries in the database that contain the same location name, and the occurrences of the name in the other texts can be automatically annotated as being "location" names.

In ongoing work we explore the usage of clustering methods to discover *hidden* events that are not explicitly coded or mentioned, but that do explain certain groupings of objects. An example in the natural history domain is the concept of *expedition*, which is not coded in the database, but which is the cause for certain animals to be found within a concentrated

period of time in a relatively restricted area by the same person or group of persons. Adding such entities and linking them to the objects that are associated to them, makes the object database searchable in more than the original dimensions.

Bibliography

Bagga, A., and B. Baldwin. "Entity-based Cross-document Coreferencing Using the Vector Space Model." *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics.* 1998. 79-85.

Bontcheva, K.., D. Maynard, H. Cunningham, and . "Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content." *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries.* 2002. 613-625.

Chapman, A.D. " Principles and Methods of Data Cleaning - Primary Species and Species Occurrence Data." *Report for the Global Biodiversity Information Facility.* Copenhagen, 2005.

Hawkins, D.M. *Identification of Outliers.* London: Chapmal and Hall, 1980.

Knorr, E. M., and R. T. Ng. "Algorithms for Mining Distance-based Outliers in Large Datasets." *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB'98).* 1998.

Sporleder, Caroline, Marieke van Erp, T. Porcelijn, and Antal van den Bosch. "Spotting the 'Odd-One-Out': Data-Driven Error Detection and Correction in Textual Databases." *Proceedings of the EACL 2006 Workshop on Adaptive Text Extraction and Mining (ATEM-06).* Trento, Italy, 2006a.

Sporleder, Caroline, Marieke van Erp, T. Porcelijn, and Antal van den Bosch. "Correcting 'Wrong-Column" Errors in Text Databases." *Proceedings of the Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn-06).* Ghent, Belgium, 2006b.

Sporleder, Caroline, Marieke van Erp, T. Porcelijn, Antal van den Bosch, and P. Arntzen. "Identifying Named Entities in Text Databases from the Natural History Domain." *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06).* Genoa, Italy, 2006c.

van Erp, M. "Bootstrapping Multilingual Geographical Gazetteers from Corpora." *Proceedings of the 11th ESSLLI Student Session.* Malaga, Spain, 2006.

Through the Reading Glass: Generating an Editorial Microcosm Through Experimental Modelling

Ron Van den Branden
(ron.vandenbranden@kantl.be)
Centre for Scholarly Editing and Document Studies
Royal Academy of Dutch Language and Literature
Edward Vanhoutte *(edward.vanhoutte@kantl.be)*
Centre for Scholarly Editing and Document Studies
Royal Academy of Dutch Language and Literature

Introduction

The Centre for Scholarly Editing and Document Studies (CTB) is preparing a digital edition of *De Trein der Traagheid*, a novella by the 20th century Flemish author Johan Daisne. The project initially aimed at a print reading edition, involving the constitution of a reading text based on a text-critical analysis of 19 witnesses of the novella's print history. However absent from the original project proposal, the TEI markup scheme was adopted early in the project as the means for digitally representing the edition. Its provisions for marking up textual variation with the so-called parallel-segmentation method informed the construction of a single XML source text containing the transcriptions of all 19 text witnesses under consideration as well as the constituted reading text, records of their mutual textual variation, and editorial annotations. The subsequent development of the electronic edition proved this unitary source text's potential for modelling a microcosm of user-generated editions. This paper will focus on characteristics, difficulties, and theoretical challenges of this particular editorial constellation, as well as the tools developed for probing into it.

Modelling

Refocusing the goal towards an electronic edition added an experimental dimension to the project. The lack of

well-established models for (creating) electronic scholarly editions forced us to conceptualise the boundaries of this particular electronic edition in the course of its development. Initially, traditional notions of scholarly editing (explicitly formalised as 'a print reading edition' in the initial project proposal) provided a good starting point for the development process. At first, development was guided by mimicking the familiar print edition model, aimed at generating a reading text with apparatus variorum from the XML source text. However, this denotative model *of* the print edition (McCarty, 2004) soon evolved to a guiding principle itself for conceptualising new ideas, an exemplary model *for* electronic editions. The added potential of an interactive edition framework allowing for user-driven input opened up new ways of exploring possible engagements of the user with the textual tradition. On a theoretical level, this exemplary model for electronic scholarly editions informed some challenging insights and rethinking of the nature of this model's object (the edition).

Technology and tools

On the most basic level, the seminal potential of the XML source text for our edition could be realised through the use of several open source XML-related technologies and tools that are currently being adopted as a standard amalgam for accessing XML resources. Key technologies for deploying XML texts like the Extensible Stylesheet Transformation Language (XSLT) and XML Query language (XQuery) allow for flexible manipulation and retrieval of XML encoded information, commonly achieved through dedicated XSLT and XQuery processors, and native XML databases.The advent of XML publishing environments like the Cocoon web development framework has made it possible to integrate these functionalities in dynamic user interfaces for presenting and querying XML content via easily accessible delivery technologies such as a web browser. This integrative potential stimulated the development of our XML text processing scripts initially developed as a tentative instrument for a specific task, to what we named the 'Morkel system', a generalisable suite of XSLT and XQuery scripts for driving electronic scholarly editions in an open source software environment.

Views on textual tradition

In the course of its experimental development the Morkel system became a tool facilitating a multi-faceted user-driven view on the textual tradition captured in the unitary XML source text.The scope of this view can be adjusted from micro- to macro-level. Users can have access to singular texts in the tradition, by requesting specific versions of the text as orientation version which presents itself as a faithful reconstruction of this text version. A broader view on the tradition can be accomplished by selecting a parallel edition, in which different episodes in the textual tradition can be viewed and contrasted, literally next to each other. This parallel presentation mode of different text versions for visual comparison resembles that of the *Versioning Machine*, developed by the Maryland Institute for Technology in the Humanities. Finally, the entire textual tradition can be taken into account when a variorum edition is selected. In its ability to compare any number of text versions with an orientation version, this variorum edition is similar to the *Juxta* tool, developed by the Applied Research in Patacriticism group at the University of Virginia. The focal point of this text comparison in the Morkel system is the contextual external apparatus variorum containing only the relevant variants for the selected comparison set and providing a locus for reorienting the edition. This scope on the textual tradition can be further refined on an intra-textual level. Complete text versions can be compared, as well as separate text divisions (one of the 33 chapters or the dedication). Where applicable when comparing different text versions, an entry to a generated apparatus is provided both at chapter level and at paragraph level.

Edition formats

One end of the delivery spectrum features the dynamic XHTML visualisation discussed so far. The versatility of XML equally allows for the generation of a PDF visualisation of the (different) edition(s), closely resembling a traditional print view of the textual tradition. A PDF rendering consists of an orientation version, either as an integral text or as a chapter sample, possibly compared to any number of comparison versions, as reflected in an inline contextualised apparatus variorum. However dynamic this generative edition frame is (Vanhoutte & Van den Branden, forthcoming), its boundaries are still present. To cater for this limitation and to enhance scientific independence, the other extreme of the delivery range is offered as well: the Morkel system equally allows users to generate pure XML renderings of the selected comparison version texts or chapters, containing their parallel-segmented inline record of the textual variation. These source texts can then be used in completely different usage scenarios, perhaps featuring completely different software environments.

Challenges

In short, the Morkel system enables users to generate their own edition(s) along 3 axes (comparison set (19 text witnesses and 1 reading text), textual scope (all or 1 of the 34 text divisions), delivery format (3 possible formats)), combining to 58 different visualisation parameters. This generates the

potential for 53.477.376 different views on the text, and problematises some traditional text theoretical concepts, as well as the defining role of tools for the electronic editions they facilitate or constitute. An obvious consequence of this generative edition paradigm (Vanhoutte & Van den Branden, forthcoming) is the promotion of each text witness to a candidate orientation version, instead of the adoption of one text version as a base text for the edition against which all other versions are calibrated. Instead, this calibration itself is made relative by the possibility of restoring each different textual witness as an autonomous landmark in the textual tradition, thence allowing a forward or backward look into the tradition. As a matter of fact, the constituted reading text itself has become integrated as 'just' a (commented) view on the textual tradition, against which all variant versions of the text can be plotted. A dynamic selection of a comparison set not only transforms the apparatus variorum to a dynamic, contextualised rendering of the relevant textual variation, but equally promotes it to a performative instrument for reorienting the edition to another point in the textual history. Due to the dynamic selection of comparison sets, the notion of variable classification becomes relativised. Discerning different types of textual variants becomes irrelevant: a variant can hold as a spelling variant in one comparison set but can change classes and become a semantic variant when compared to another version in the textual tradition. The search capabilities of the Morkel system even extend the view on the textual tradition from text level, by allowing simple search operations inside one text version (intra-textual), to collection level, by allowing complex search operations over different text versions (extra-textual). To conclude, this generative paradigm for electronic scholarly editions seems to articulate the defining role of the specific tools for accessing electronic texts more sharply. On its own, the XML representation of text-critical research is a valuable record of scientific labour, but it is the specific (generative) interface which instantiates it as an editorial microcosm by providing a range of user-driven access methods that enable dynamic exploration of a textual tradition. The characteristics and exact nature of this user-driven scholarly edition or constellation of editions is strongly determined by the boundaries this generative interface provides, allowing for a microscopic, telescopic, stereoscopic, or kaleidoscopic view on the textual tradition.

Bibliography

Juxta . <http://www.patacriticism.org/juxta />

McCarty, Willard. "Modeling: A Study in Words and Meanings." *A Companion to the Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing Ltd, 2004. 254-70.

Sperberg-McQueen, C. M., and Lou Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange, XML-compatible edition.* Oxford: TEI Consortium, 2002. <http://www.tei-c.org/P4X/>

Van Hulle, Dirk . *Textual Awareness: A Genetic Study of Late Works by Joyce, Proust & Mann.* Ann Arbor, MI: University of Michigan Press, 2004.

Vanhoutte, Edward, and Ron Van den Branden. "Describing, Transcribing, Encoding, and Editing Modern Correspondence Material: A Textbase Approach." *Computing the Edition.* Toronto: Toronto University Press, Forthcoming.

Versioning Machine . <http://www.v-machine.org />

TEI By Example

Ron Van den Branden
(ron.vandenbranden@kantl.be)
Centre for Scholarly Editing and Document Studies
Royal Academy of Dutch Language and Literature
***Edward Vanhoutte** (edward.vanhoutte@kantl.be)*
Centre for Scholarly Editing and Document Studies
Royal Academy of Dutch Language and Literature
***Melissa Terras** (m.terras@ucl.ac.uk)*
University College London

The TEI (Text Encoding Initiative)[1] has provided a complex and comprehensive system of provisions for scholarly text encoding. Although a major focus of the "digital humanities" domain, there is little evidence that it has been embraced by and taught to humanities students at university level, which is important to encourage the adoption of the TEI's recommendations and the widespread use of its text encoding guidelines. Very little training material exists for either individuals who wish to teach themselves TEI, or university level teachers who wish to have access to adequate training materials to use with their classes.

If the digital humanities community wants to promote the TEI markup framework as a serious candidate for dedicated courses in the booming curricula on digital humanities, humanities computing, digital culture, or humanities informatics, to name just a few of the labels this archipelago of disciplines gets, as well as integrating the TEI further with digital librarianship, then there is an urgent need for an on-line TEI course by example which is less generic than the two tutorials published on the TEI website (Sperberg-McQueen and Burnard (2002a) Sperberg-McQueen and Burnard (2002b)) and more user friendly, comprehensive, and up to date than the "latest" Teach Yourself TEI materials currently online (all of which date from 2002).[2]

The *TEI By Example*[3] project is currently developing a range of freely available online tutorials walking individuals through the different stages in marking up a document in TEI. *TEI By Example* aims to provide online tutorials which give examples for users of all levels. Examples will be provided of different document types, with varying degrees in the granularity of markup, to provide a useful teaching and reference aid for those involved in the marking up of texts. Likewise, the availability of a software toolkit for teaching text encoding will support the potential trainers to take up the challenge to teach TEI on several occasions. In order to support multilingualism in the text encoding community, the on-line tutorials are being considered for translation into a number of languages. The translations proper, however, are outside the scope of this project.

The deliverables of the project are: on-line tutorials *TEI By Example*, a printable PDF version of the tutorials *TEI By Example*, an on-line software toolkit for text encoding, a downloadable CD-ROM image for burning off-line toolkits for use by course participants, and adequate documentation to enable the tutorials to be used elsewhere if needed.

Project partners are The Centre for Scholarly Editing and Document Studies (CTB)[4] of the Royal Academy of Dutch Language and Literature, the Centre for Computing in the Humanities (CCH)[5] of King's College London, and the School for Library, Archive, and Information Studies (SLAIS)[6] of University College London, with an international advisory board consisting of experts in textual encoding and markup. The deliverables will be published and hosted by CCH (King's College London) under endorsement by the Association of Literary and Linguistic Computing. Development of the tutorials began in October 2006 and will start to appear online in Spring 2007. It is hoped that the project results will be relevant to the trainers of TEI, the students of TEI, the text encoding, and the humanities computing community in general.

A major point of attention at the start of the project was the status of the TEI model. Since early 2002, the TEI Consortium has been engaged in a major (backward-incompatible) revision of the TEI specification, migrating it from version P4 (released in 2002, see TEI (2004) to P5 (2005 onwards, see TEI (2005)). Featuring more than just changes in the markup model and the content of the guidelines, P5 entails an overhaul of the complete production process of the standard. It seems that the timing of this *TEI By Example* project coincides with a turning point in the transition of TEI P4 to P5: the advantages of P5 adoption for this project seemed to outweigh the disadvantages of P4. The most recent snapshot indeed suggests that stability is at hand (Van den Branden 2006). As a result, the project is developing materials in P5.

Eight tutorials are under construction. The first, an introduction to text encoding and the TEI, encourages the user to explore textual encoding and markup to foster an understanding of why this is useful, or even necessary, to allow texts to be processed automatically and used and understood by others. The TEI header tutorial covers the type of information and metadata captured in the header element. Three tutorials focus on examples of individual types of text: Prose, Poetry, and Drama, and a further two tutorials deal with examples of Manuscript Transcription and Scholarly Editing. The final tutorial

investigates how the TEI can be customized, and the use of ODD and Roma.

Although under development at time of writing of this poster proposal, the tutorials will be available for demonstration by Digital Humanities 2007, and this poster aims to highlight the format, structure, testing, and development of the tutorials. As well as informing the academic audience at DH2007, and publicising the project, this poster will aim for feedback to allow further development of the online material.

---

1. <http://www.tei-c.org/>
2. <ttp://www.tei-c.org/Tutorials/index-latest.html>
3. <http://www.teibyexample.org>
4. <http://www.kantl.be/ctb/>
5. <http://www.kcl.ac.uk/schools/humanities/cch/>
6. <http://www.slais.ucl.ac.uk/>

Bibliography

Burnard, Lou, and Syd Bauman, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* TEI Consortium, 2005. <http://www.tei-c.org.uk/P5/Guidelines/index.html>

Sperberg-McQueen, C. M., and Lou Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange, XML Compatible Edition.* Oxford: TEI Consortium, 2002. <http://www.tei-c.org/P4X/index.html>

Sperberg-McQueen, C. M., and Lou Burnard. "A Gentle Introduction to XML." *TEI P4: Guidelines for Electronic Text Encoding and Interchange, XML Compatible Edition.* Ed. C. M. Sperberg-McQueen and Lou Burnard. Oxford: TEI Consortium, 2002a. <http://www.tei-c.org/P4X/SG.html>

Sperberg-McQueen, C. M., and Lou Burnard. *TEI Lite: An introduction to Text Encoding for Interchange.* 2002b. <http://www.tei-c.org/Lite/teiu5_en.html>

Van den Branden, Ron. *[TBE-R001] - TEI by Example, Initial Report, 2006/06/09.* 2006. <http://www.kantl.be/ctb/project/2006/TBE-R001.htm>

The Complete Works of W.F. Hermans. Using Automatic Text Comparison and XML for a Voluminous Edition

Bert Van Elsacker *(bertve@gmail.com)*

Huygens Instituut
The Netherlands

Introduction

In November 2005 the first volume of the Volledige Werken (Complete Works) of Willem Frederik Hermans appeared. This publication marked the official beginning of the largest Dutch edition project ever undertaken in the field of modern literature. There are two sides to the edition: a publication in print, aimed at a general public, and a web site, where the reader can find the editorial principles, a description of the textual history and listings of editorial emendations. The project is exceptional not only because of its size, but also because right from the beginning it has been set up as an experimental digital research project.

The initial impetus came from the need for automated text comparison. An academic edition cannot do without careful comparison of the different versions of a text. In the case of Hermans, about one third of all print editions has been revised, which brings the volume of research material to about 50,000 pages. This implied manual collation was not feasible. We will give an overview of the procedure to transform a large amount of print material into accurate digital text, demonstrate a system which outputs XML-TEI-encoded collation results and show an example of the possible uses of these documents.

The edition

Willem Frederik Hermans is widely regarded as the most important Dutch author of the second half of the twentieth century. In addition to novels, Hermans (1921-1995) wrote short stories, plays, poetry and essays; he also translated several texts, including Ludwig Wittgenstein's Tractatus Logico-Philosophicus. His work is receiving more and more international attention; over the past few years several novels have been translated into German, French and English. Beyond Sleep (Nooit meer slapen) was hailed in the English press as a

forgotten masterpiece of post-war European literature, or in the words of the Times ‘a welcome if belated introduction to an original and challenging voice in modern European literature’. Hermans was a very prolific author. The edition will consist of a total of 24 volumes, each with an average of about 800 pages. The Huygens Institute (a research institute within the Royal Netherlands Academy of Arts and Sciences) is responsible for the preparation of a reliable edition text in accordance to scholarly standards. For each work, the last version Hermans authorised is the point of departure for the critical edition. The editors review this version in the light of the results of archival research and of a comparison with selected other versions of the text.

Automatic text comparison

Traditionally the meticulous comparison of selected versions of a text has been one of the most important tasks of an editor. Until a few decades ago manual collation – which is extremely time-consuming – was the only possibility. However, in computer science theoretical and practical research into automated text comparison has been taking place since the 1970s. In computer science, the general task of "text comparison" has to be expressed as a formal procedure in order to make the development of an algorithm possible. More precisely, "text comparison" has been understood as a procedure which results in a list of changes between two versions of a text. The application of all changes to version A transforms version A into version B. The algorithm should try to keep the list as short as possible. This approach has led to a basic algorithm by Eugene Myers, on which some variations have been developed, and various implementations, of which the Unix/Linux tool ‘diff’ is the most widely used. Source code implementing this algorithm is readily available for all major programming languages.

Somewhat apart from these developments in computer science, in the world of scholarly editing there have also been initiatives to use computers for text comparison and, by extension, for the production of editions. The best-known examples are Peter Robinson’s program Collate and Wilhelm Ott’s TUSTEP. Collate is only available for the Macintosh Classic platform, an operating system which has now been replaced by OS X. The program is particularly suitable for older texts which have been divided into relatively short passages beforehand and in which there are not too many long or complex variants. The algorithm used to extract variants remains undocumented. Mr. Robinson has announced a successor to Collate called EDITION. TUSTEP is actually a comprehensive environment for textual research and the production of editions. It is a rather complex instrument to work with and is more something like a programming language in itself. Another impediment is the absence of a graphical user interface (GUI). Recently, another interesting application, called Juxta, has become available. Unfortunately, for now the program lacks the option to export the collation results. According to the project web site, this will change with the next release later this year.

On the one hand automatic text comparison has great advantages: the comparison is based on a formally defined algorithm, free of errors and in principle reproducible by others. Moreover, the use of computers saves a huge amount of time, which may be of crucial importance as often resources are lacking to collate texts manually. On the other hand, there is no ready-made solution for the average user; whichever option is chosen, some extra training of the prospective user is necessary, and experience with scripting will probably come in handy. Of course this learning process also takes time and energy. For small projects the effort may outweigh the advantages, but considering the size of the Hermans project, in this case the investment did seem worthwhile. To date we have made extensive use of Collate and software tools such as diff.

Automatic collation requires accurate digital sources. For a scholarly edition, OCR-results aren't good enough. A checking system is necessary. Moreover, some presentational markup like quotes, the use of bold, italic etc., and expressive white space has to be captured as semantic markup, while other presentational features (page width, fonts used...) have no importance, in any case for the Hermans edition, and can be quietly disregarded. So a phase of checking and automatic encoding before the actual collation is required. We will discuss this operation in more detail in our poster presentation.

XML-encoded collation results

The result of these preparations are reliable and detailed overviews of all the differences between sometimes a dozen versions of a text, encoded as a base text with in-line apparatus conforming to the TEI Guidelines. This means that a complex text history can be examined down to the tiniest detail in a manageable way. The use of XML enables the editor to observe patterns in variants, to categorize findings, and to examine them in greater detail. Due to systematic encoding of the material the accumulated data can be searched accurately and checked for correlations, and working hypotheses can be continually tested (for example by using XML search languages such as XPath and XQuery) and if necessary modified.

Currently, we are working on ways to present the multiple versions of the text and the conclusions reached in research in a dynamic way, partly on the basis of the digital research documentation and the findings of the analysis of De tranen der acacia’s (a major novel which appeared in the first volume). During the poster session, we will display the short story

'Paranoia'. The digital presentation of this story contains the reading text of the Hermans edition and all the text versions which were of importance for the editorial research. We intend to place a short introductory section before the full-text documentation in which some important revisions in the short story are discussed, as a reader's guide. For example, in the version of this story in the first publication in book form in 1953 Hermans puts more emphasis on the theme of the housing shortage, which was a key concern in the Netherlands after the war, especially in Amsterdam, where the story is set. Due to adjustments in substance and narrative, the events in the book version of 'Paranoia' are described more from the perspective of the character Cleever than in the magazine publication of 1948, a significant revision in a story about someone who suffers from persecution mania.

In the digital publication we will integrate observations and analyses into the texts to which they refer, as a form of empirical evidence. Relevant text passages will therefore be tagged in the online presentation so that they can be seen separately and in context. We are also examining other presentation options. There are analyses conceivable, such as a narratological study of narrative structure, which in the form of hypertext can serve as a point of access or a guide to Hermans's work. Ideally, in future digital text presentations text and research will constitute an integrated collection of data which can constantly be consulted, modified and expanded. Or as Hermans himself once put it: '...a collection, an enormous accumulation of movements and ideas.'

Bibliography

Juxta. <http://www.patacriticism.org/juxta/>

Myers, Eugene W. "An O(ND) Difference Algorithm and Its Variations." *Algorithmica* 1.2 (1986): 251-266.

Ott, Wilhelm. "Strategies and Tools for Textual Scholarship: The Tübingen System of Text Processing Programs (TUSTEP)." *Literary & Linguistic Computing* 15.1 (2000): 93-108.

Robinson, Peter. *Collate 2: A User Guide.* Oxford: The Computers and Variant Texts Project, 1994.

SDH/SEMI Panel: Explorations in a Variety of Interfaces for the Reading of a Database

Christian Vandendorpe
(Christian.Vandendorpe@uottawa.ca)
Département de lettres françaises
Université d'Ottawa
Stan Ruecker *(sruecker@ualberta.ca)*
University of Alberta
Stéfan Sinclair *(sgsinclair@gmail.com)*
McMaster University
Dominic Forest *(dominic.forest@umontreal.ca)*
École de bibliothéconomie et des sciences de l'information
Université de Montréal

A Textual Database of Dreams and the New Context of Reading

Christian Vandendorpe (Département de letters françaises, Université d'Ottawa)

The advent of the web has fostered the development of specialised textual databases, dedicated to tabular collections of snippets of our literary heritage. Their purpose is to facilitate the study of a particular phenomenon through the centuries. Henceforth, the act of reading tends to adopt a tabular way of reading as the one advocated by Claude Levy Strauss for the study of myths. In order to focus on a specific case, we shall examine the textual mysql database <http://www.reves.ca>, a collection of some fifteen hundred dream narratives from various cultures since Homer and the Bible, and discuss its goals and its initial interface. The challenge is to design an interface that would be both efficient and exciting, inducing in the reader the desire of reading and helping him or her to explore the database and to make meaningful relations between the data.

The Mandala as an Interface to a Textual Database

Stan Ruecker (University of Alberta) and *Stéfan Sinclair* (McMaster University)

The Mandala is a Java-xml application that makes visible the relations between sets of data. Its playful interface is designed to elicit a meaningful reading activity by the user in order to facilitate an in-depth exploration of the contents. We shall examine how it works with <http://www.reves.ca>, its potential and its pertinence in the new context of reading.

Towards More Powerful Semantic Interfaces

Dominic Forest (École de bibliothéconomie et des sciences de l'information, Université de Montréal)

In this presentation, I will discuss the preliminary results of a research initiative where we explored the relevance of applying text mining techniques on a literary corpus. The corpus used in this work is composed of short thematically-related literary texts stored in a database. The main objective of this project is twofold. First, we want to explore if traditional text mining techniques can successfully be used to assist the extraction and analysis of relevant information found in literary texts. Secondly, we also want to explore if computer-assisted text analysis tools and text mining techniques can be useful to help access to valuable literary information stored in a textual database. This talk will be divided in three main parts. First, I will present the fundamental text mining concepts and techniques. In the second part, I will describe the computer-assisted text analysis processes that were applied on the online textual database (<http://www.reves.ca>). Finally, I will discuss some of the results we obtained and give an overview of the future work that will be done in this project.

A Descriptive Classification Generator for Electronic Editions

Edward Vanhoutte *(edward.vanhoutte@kantl.be)*
Centre for Scholarly Editing and Document Studies
Royal Academy of Dutch Language and Literature
Ron Van den Branden
(ron.vandenbranden@kantl.be)
Centre for Scholarly Editing and Document Studies
Royal Academy of Dutch Language and Literature

Introduction

Electronic scholarly editions which mimick conventional models of scholarly editions as prescribed by different theoretical and methodological schools provide denotative models (Geertz, 1993, p. 93) which thrive on the otherness of the digital medium but re-emphasize the computational aspect of the 'computer-based' (Steding, 2002), 'Computergestützte' (Kamzelak, 1999), or computer-assisted scholarly edition. The isomorphism between the digital and the print medium aimed at by the application of computational techniques to the praxis of scholarly editing confirms what we already know. What is interesting, however, is not the degree to which the computer can assist the editor in digitizing, creating, and publishing an edition, but the intentional artefacts which are built by using the computer as a modelling tool (Smith, 2002). They are instrumental in two crucial activities of humanities research, that is, the discovery of meaning and the making of meaning. As products of (experimental) modelling, their purpose is 'to achieve failure so as to raise and point the question of how we know what we know' (McCarty, 1999), 'what we do not know', and 'to give us what we do not yet have.' (McCarty, 2004, p. 255) This paper will address the role of experimental modelling and the assessment of exemplary models of scholarly editions in the development of a useful classification, typology, and description of electronic editions.

Classification

As psycholinguistic research has shown, categorization is innate in human cognition (Giannakopoulou, 2003) and

involves the formation and use of patterns in a self-maximizing system (deBono, 1978, pp. 25-43). Operations which can be classified under this scholarly primitive are naming, labelling, classifying, cataloguing, indexing, sorting, etc. Categorizing as a mind process can result in the production of formalized instruments such as bibliographies, indexes, catalogues, classification schemes, and taxonomies for which advanced subject analysis is needed. For the most part, however, it remains a culturally determined mind tool particularly where it is used for the selection of usefulness. This means that categorizing is not determined by how the world is, but tries to develop convenient ways in which to represent it (Hacking, 1999, p. 33).

Traditions and Typologies

Textual scholarship is fragmented by the development of different theories, methods, and praxes which are based on a diversity of attitudes and perspectives (author, language, audience, function, format, etc.). This becomes especially clear when studying current typologies and classification schemes for scholarly editions. Heinrich Meyer (1992) surveyed the literature on textual scholarship in Germany in the twentieth century and listed more than forty names for different types of editions that were used. As he argued, the 'ausgabentypoligische Terminologiewirrwarr' (Meyer, 1992, p. 17) is the result of a methodological pluralism both inside and across editorial traditions.

As a consequence, there is no one theoretical paradigm for textual scholarship across all traditions, periods, languages, and authors and there is no one universally applicable taxonomy of editorial types. Moreover, the existing taxonomies are seldom internally consistent in their applied perspective. The simplified representation in the German school, for instance, offers a taxonomy which runs from the archive edition over the historical-critical edition to the study and the reading edition (Kanzog, 1970, pp. 9-44). Where 'archive edition' denotes the archival function of this type of edition and hints at the extent of the documentary set presented, 'historical-critical' refers both to the method used to create the edition and the format in which that edition comes before the user. The study-edition and reading edition, on the other hand, identify the envisioned function of the product and its intended audience in their naming. In the Anglo-American tradition, the copy-text edition refers to a specific theory of establishing a text whereas the types of scholarly editions David Greetham mentions in his *Textual Scholarship. An Introduction* mainly refer to the format or appearance of the edition, such as 'parallel print edition', 'variorum edition', or 'type facsimile edition', or to a combination of format and method such as 'Eclectic Clear-Text Edition with Multiple Apparatus'. (Greetham, 1994, p. 383)

The least useful typology of scholarly editions is based on the publication medium. Here we have print edition, hybrid edition, and electronic edition. Especially this last one is often presented as a meaningful class while it is widely used to name almost anything which is available in an electronic format. A sad example is, for instance, the édition électronique of the correspondence of René Descartes which is nothing more than a 35 page MSWord file which has been made available online <http://classiques.uqac.ca/classiques/Descartes/correspondance/descartes_correspondance.doc>.

Electronic editions

With respect to the classification of electronic editions, it becomes difficult to maintain the application of conventional typologies and taxonomies, or ignore them altogether. The danger of a normative typology and hence a rigid theoretical frame for textual scholarship is that it establishes its principles firmly without allowing the advancement of its theories, methodologies, and practices. However, as a scholarly discipline, scholarly editing should be interested in both. Especially when, in the case of electronic scholarly editing, exemplary modeling is employed as a scholarly method to generate electronic editions rather than the epigonuous application of rigid theory and method to the electronic edition.

Classification Generator

For reasons of identification and bibliographic research on electronic editions (Lavagnino, 1996; Dahlström, 2002; Kirschenbaum, 2002; Van der Weel, 2005), there is a need for some integrated scheme by which editors of electronic editions can describe their edition according to several parameters. With the classification generator which we propose here, we believe we have developed a tool which can be of aid to that purpose.

The classification generator is an on-line tool which allows the editor to input the details of the electronic edition atomized in meta-information on the edited text (language, period, genre) and information on the edition via a user-friendly form. The latter minimally contains details about method, intended audience, content, format, encoding, technology, function, and functionality of the edition. Once the edition is described according to these parameters, a descriptive classification code is generated that can be included in the published edition. This classification code is an alphanumerical string that exactly describes the electronic edition from multiple perspectives. The classification generator serves at least three goals. First, it liberates the field of electronic scholarly editing from the conventional text-editorial theories with their rigid and

inconsistent prescriptive typologies. Instead the classification generator atomizes the different facets of the electronic edition and presents the sum total of this documentation as a description of the product. Second, the user confronted with an electronic edition gets a detailed description of the kind of electronic edition one is using on inputting the classification code in the classification generator. Third, the codes derived from the classification generator can be of use for an (analytical) bibliography of electronic editions. The description of an improved re-release of an electronic edition will generate a different classification code which could be collated against the codes of other releases of the same edition.

A last feature of the classification generator is the option to register an edition's classification code together with a formal bibliographic description in a database. This database will allow theorists of electronic scholarship and bibliographers of new media to perform interesting forms of analysis on its contents.

Bibliography

Dahlström, Mats. "Nya medier, gamla verktyg." *Human IT* 6.4 (): 71-116. <http://www.hb.se/bhs/ith/4-02/md.pdf>

deBono, Edward . *Lateral Thinking: A Textbook of Creativity.* Harmondsworth: Peguin, 1978.

Geertz, Clifford. *The Interpretation of Cultures.* London: Fontana, 1993.

Giannakopoulou, Anastasia. "Prototype Theory: An Evaluation." *Ecloga Online Journal* 3 (2003). <http://www.strath.ac.uk/ecloga/Giannakopoulou.htm>

Greetham, D. C. *Textual Scholarship. An Introduction.* New York: Garland Publishing, 1994.

Hacking, Ian. *The Social Construction of What?*Cambridge, MA: Harvard University Press, 1999.

Kamzelak, Roland, ed. *Computergestützte Text-Edition.* Beihefte zu Edition, 12. Tübingen: Niemeyer, 1999.

Kanzog, Klaus . *Prolegomena zu einer historisch-kritischen Ausgabe der Werke Heinrich von Kleists. Theorie und Praxis einer modernen Klassiker-Edition.* München: Carl Hanser Verlag, 1970.

Kirschenbaum, Matthew G. "Editing the Interface: Textual Studies and First Generation Electronic Objects." *TEXT: An Interdisciplinary Annual of Textual Studies* 14 (2002): 15-51.

Lavagnino, John . "The Analytical Bibliography of Electronic Texts." Paper presented at ALLC/ACH 1996, Bergen, Norway, June 25, 1996. 1996.

McCarty, Willard. "Humanities Computing as Interdiscipline." A seminar in the series "Is Humanities Computing an Academic Discipline?", held under the auspices of the Institute for Advanced Technology in the Humanities (IATH), at the University of Virginia, Guy Fawkes Day 1999. 1999. <http://ilex.cc.kcl.ac.uk/wlm/essays/inter/><http://www.iath.virginia.edu/hcs/mccarty.html>

McCarty, Willard. "Modeling: A Study in Words and Meanings." *A Companion to the Digital Humanities.* Ed. Susan Schreibman, Ray Siemens and John Unsworth. Oxford: Blackwell Publishing Ltd, 2004. 254-70.

Smith, Brian Cantwell. "The Foundations of Computing." *Computationalism: New Directions.* Ed. Matthias Scheutz. Cambridge, MA: MIT Press, 2002. <http://www.jhfc.duke.edu/jenkins/courses/p256/readings/p256foc.pdf>

Steding, Sören A. *Computer-Based Scholarly Editions: Context - Concept - Creation - Clientele.* Berlin: Logos Verlag, 2002.

Van der Weel, Adriaan. "Bibliography for the New Media." *Qaerendo* 35.1-2 (2005).

MusicXML: An XML Based Approach to Automatic Musicological Analysis

Raffaele Viglianti *(raffaeleviglianti@gmail.com)*
University of Pisa
Italy

1. Introduction

Computational music is a young discipline undergoing ever increasing development. Most research conducted to date concerns digital audio, multimedia, automatic composition, score writing, etc. A humanities-oriented approach (i.e. musicological) is a more recent, less well-defined development, though it is also growing.

The application of such studies regard, to mention only a few:

- automatic extraction of statistics;
- verification of grammar rules applied to music;
- automatic application of models for formal and harmonic analysis;
- validation of models for different analysis types.

The MIDI format is the technology most often used in computational music. Developed in the early 1980s, MIDI is a digital representation of music, used especially for communication between electronic musical instruments, including computers. The MIDI message associates an integer number to each note (e.g. central C is 60), so it is suitable to a wide range of studies.

A newer technology, MusicXML, offers the same possibilities and more, but requires different techniques for data retrieval.

The following abstract presents the results of a statistical analysis of the tenor parts of some operas by Giacomo Puccini, encoded in MusicXML for the Research Institution "Centro Studi Giacomo Puccini" in Lucca, Italy. Specifically, it examines the arias of Rodolfo from *La Bohème* and Principe Calaf from *Turandot.*

The encoding of these musical pieces was discussed in one of the author's thesis dissertation for his degree in Computer Science for the Humanities at the University of Pisa, under the supervision of Professor Elena Pierazzo. The dissertation deals with the possibilities for automatic applications illustrated in section 4 herein.

2. What is MusicXML?

LLC is an American Internet music publishing and software company; it developed and promoted the MusicXML encoding language, which can represent the music notation system established in the West since the 17th Century.

MusicXML is a royalty-free format that implements all the features offered by XML technology:

- data structure
- modularity
- extensibilty
- possibility of querying and interaction through XML family technologies.

MusicXML is designed to represent a score in a digital format, so most of information contained in a human-readable score is retrievable in the encoding and can be used for various kinds of analyses. Features such as lyrics, author or reviewer notes, dynamics, agogic indications, and so on, are encoded with their position in the score.

An important part of the encoding is the Score Header, which contains metadata, such as author, poet, work number, instruments, etc. An instrument data section even enables specifying the MIDI instrument channel to be used to play the music.

3. Automatic statistical extraction: an example application.

The encoded tenor parts have been inserted into eXist, an Open Source XML database, to be queried with the XQuery language.

The goal was to extract statistical data with relevance to the humanities. Indeed, such data was used for a musicological dissertation on vocalism in Puccini's works.

The statistics are based on the criteria proposed by Marco Gilardone and Franco Fussi in *Le voci di Puccini. Un'indagine sul canto* (The Voices of Puccini. An inquiry on singing). The statistics presented in such work were extended and adapted to include the *ossias*.[1]

The statistics are the percentage occurrences in the distribution of notes according to the following classification of tones:

Table 1 classification of tenor tones

ACUTE TONES	Upper limit
	f3 – f# 3
MIDDLE TONES	E3 A2
	G2 – G#2
GRAVE TONES	Lower limit

The upper and lower limits are the extreme notes of the analyzed part; the numbers beside to the notes denote the octave; the notes highlighted in grey are transition tones, which are interesting due to their difficulty of execution within the part.

The original statistics were extended to account for the incidence of dynamic aspects such as *piano, pianissimo, mezzoforte, forte* in transition tones and acute tones.

The XQuery query language was used, as it offers many possibilities for data extraction from an XML database, though it is often redundant and does not provide for complex structures such as arrays or the like.

Starting with the retrieved numerical data, a musicologist can construct a dramatic character profile. For example, Calaf's singing parts exhibit a high percentage of acute tones (27.39%), transition tones (17.63%) and a high ratio between acute and grave tones. This is a deliberate aesthetic choice, made to portray a character in less realistic terms than other tenor roles, perhaps because he is part of a fantastic drama; indeed, his vocality tends to be heightened.

Rodolfo, instead, exhibits 'milder' musical behavior; he sings less transition notes (10.37%), and those with dynamic indications are sung in piano. The middle tones are greatly predominant (69.63%); Rodolfo is indeed a character who tends to be more intimate, and displays less dramatic presence.

All the statistics are produced via a query, and each piece of information needed for calculations is furnished directly from the encoded score.

Beginning with this model, many other searches may be carried out, such as graphs or division of the notes in each act, etc. Deeper analysis with the model provides for clearer delineation of a character's profile.

4. Thematic extraction based on Retì's musicological analysis

In the 1950s, Rudolf Retì developed a musicological analysis based on the repetition of thematic cells. Such cells may occur in different ways and can be found by considering the melodic interval.[2]

Consider, for example, the following sequence of notes:

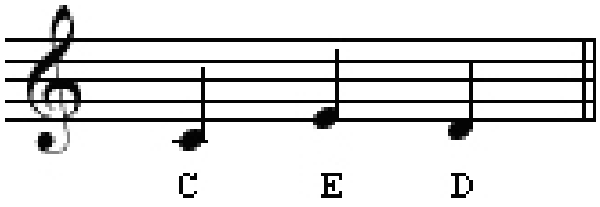


Figure 1

Between C and E there is an interval of two tones, while E and D are separated by only one tone. If the following sequence

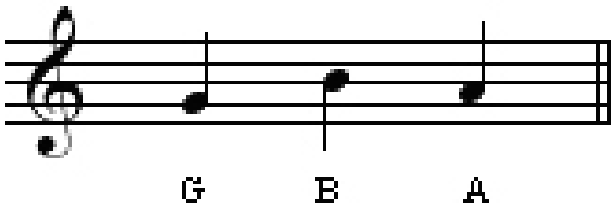


Figure 2

occurs later on in the part, it can be compared to the sequence in figure 1, because the intervals correspond: there is an interval of two tones between G and B, and one tone between B and A.

Retì views entire compositions as based on a few thematic cells. The cell can occur in the same sequence as the initial one or in a different sequence, that is:

- notes in inverse order
- exchanged position of the notes
- transposition to different notes, but with the same interval.

Any transformation of the cell can also occur with "interpolated" notes, which are inserted between the principal notes.

Starting from the corpus encoded in the author's dissertation for the Research Institute Centro Puccini, an algorithm was envisioned for automatic retrieval of the repercussions of such thematic cells.

Although such algorithm was implemented with XQuery, it proved unable to fulfill the task, because of the lack of advanced structures in XQuery. The study is still ongoing, as is the implementation of the algorithm in other programming languages, in particular, Java or Perl.

5. Limits of MusicXML

Even if MusicXML offers many possibilities for research, not every prerequisite of high level academic research can be satisfied: it is designed to be a sufficient but not optimal representation of musical scores. For instance, with MusicXML it is impossible to encode ossias, but for this research it was enough to encode only one of the two options of the score. On the other hand , the encoding model would not be optimal for a study that has as a goal to represent a score exactly as it appears. Anyway MusicXML is extensible and many other encoding models can be developed starting from the actual model. At this stage it should be pointed out that there are other encoding models that are starting to be used in the academic research, such as MEI (Music Encoding Initiative). MEI is designed to represent with effectiveness also information about the physical support and provides a more complex structure of metadata. Still the weakness of MEI is the lack of plug-ins that can convert from score writing software or from optical recognition programs. As a matter of fact, most research based on MEI uses XSLT technology to convert from MusicXML to MEI. This area of research is constantly evolving.

6. Conclusions

The abstract presents MusicXML as an alternative technology for conducting automatic analysis of music, adopting a statistical approach. The potentials of such techniques are considerable: in the future they are likely to be applied for advanced research, such as pattern matching and automatic formal analysis, or for added encoding, including commentaries, metadata for retrieval, harmonic analyses, as well as for retrieval from digital collections of scores.

1. An *ossìa* is an alternative way to execute part of a measure or one or more measures.
2. A melodic interval is the tonal distance between two notes in sequence.

The Use of TEI and OAI in Manuscript, an Informational-Analytical System

Pavel A. Votincev *(pvl@udsu.ru)*
Udmurtia State University, Izhevsk, Russia

Currently there are at least a few hundred ancient Slavic manuscripts of the 11th to 14th centuries, which, despite their significant scholarly, cultural, and historic value, are not accessible to a wide range of researchers of various specialties and to all who are interested in Russian history and culture. The reason is simple: a large portion of the manuscripts are still unpublished, and if they are published, these editions are rareties, accessible only in major research libraries, which are often located far from scholars' places of work. Even publishing a manuscript often requires constant consultation of the original, especially if studying its paleography, appearance, or other similar questions. Furthermore, existing print editions unfortunately usually lack reference material necessary for getting acquainted with a text and studying it, and the manuscripts themselves are published with poor quality. The solution to the inaccessibility of ancient and medieval manuscripts for research, teaching, and popular use is the creation of electronic publications capable not only of replacing print editions but also of exceeding their capabilities: (a) by providing continual access to digital copies of manuscripts and reference material on the Internet; (b) by providing the user with the capability of retrieving data from a multifunctional interface not only on the level of meta- and analytical data but also on the level of grammatical, semantic, and thematic characteristics of texts; and (c) by allowing collaborative work on a manuscript.

One example of an information system designed to solve some of the above problems is Manuscript, an informational-analytical system.

Manuscript is oriented toward the entry, storage, automatic manipulation, research, and publication of ancient texts having an identical graphico-orthographic appearance and structure to the original. Currently Manuscript contains a few dozen unique ancient and medieval Slavic texts.

Currently Manuscript consists of a few multifunctional modules that interact with a single database containing a hierarchically structured data:

- a specialized editor,
- a query and retrieval module,
- a module for electronic publishing,
- a module of linguistic dictionaries, and
- a module for loading texts and reference material for print publication.

The main task right now of the team behind Manuscript is to present the accumulated information for analysis and study by a wide range of researchers. The main path: to create specialized modules intended for muti-function, user-friendly, and straightforward manipulation of the data. The second task is to create the means to exchange meta-, analytical, and textual data between teams of researchers creating analogous collections of material, with the goal of presenting users with the ability to use manuscript- and text-specific tools for working with ancient documents, created by various teams.

It's essential to use standard formats and protocols to maximize access to the various tools for aggregating information in Manuscript.

The Open Archive Initiative (OAI) and Text Encoding Initiative (TEI) were chosen as the basis for solving this task:

- OAI – for indexing the existing texts and their fragments;
- TEI – for presenting meta- and analytical information, and likewise the texts themselves.

Manuscript is currently working to become an OAI data provider. A new module is being developed to present the metadata in Dublin Core format. This module will allow the gathering of metadata not only from the texts themselves but also from their fragments (various elements of the hierarchy).

Users can retrieve metadata or full text from Manuscript in TEI XML format using a unique identifier.

The TEI Guidelines are well-established around the world, so it was entirely logical for the developers of Manuscript to use it for the metadata and full text of the manuscripts.

Manuscript uses a network model of data for describing the objects (entities) of manuscripts and the relations between them, so the main problem with using TEI was XML's requirement of a strict hierarchy. One consequence of this is that it's impossible to receive all information about a text in Manuscript in one document.

Two modules are being developed to integrate the technology of Manuscript with the TEI.

The module for storage and manipulation of meta- and analytical data is intended for working with archeographical, textological, and other findings from the manuscripts, texts, and fragments themselves, and also for formulating queries and retrieving information using these findings. Searching in this module is possible only using meta- and analytical data.

A special module processes user queries for retrieving certain elements of text in TEI format from Manuscript. The main capabilities of the module are the following:

- retrieving of any element of text, stored in Manuscript as an independent unit, in TEI format;
- setting priorities in the creation of a TEI document: This option is necessary in order to get around the single-hierarchy limitation of XML. The user can rank the priority of various structures within a document;
- receiving meta- and analytical information with or without full text;
- The queried document is created at the moment of query, so any change will be available immediately.

Manuscript and its tools are being continually developed. During the presentation, Manuscript, including the above modules and technologies, will be demonstrated.

Translated from the Russian by Kevin S. Hawkins.

Interoperability of Metadata for Thematic Research Collections: A Model Based on the *Walt Whitman Archive*

Katherine L. Walter *(kwalter1@unl.edu)*
Center for Digital Research in the Humanities
University of Nebraska - Lincoln
Brett Barney *(bbarney2@unl.edu)*
Walt Whitman Archive
University of Nebraska - Lincoln
Julia Flanders *(Julia_Flanders@brown.edu)*
Women Writer's Project
Brown University
Terence Catapano *(thc4@columbia.edu)*
Columbia University Libraries Digital Program
Daniel Pitti *(dpitti@Virginia.edu)*
Institute for Advanced Technology in the Humanities
University of Virginia

Topic:

Created by scholars in collaboration with librarians/archivists, thematic research collections are directed primarily at other scholars, though they are also used by students and by the general public. Their development and preservation pose several challenges: 1) they require very detailed encoding and metadata to serve their demanding audience; 2) they often involve materials which are dispersed; and, 3) critically, they incorporate a variety of different types of materials with different digitization and metadata requirements. Digital thematic research collections often include texts, images, finding aids to archival material, and administrative records. While standards have been developed for each of these data and metadata types (e.g., TEI, TIFF, EAD, and MODS), there has not yet been a disciplined effort to integrate the standards and explore where overlap and commonalities are, and to document best practices for recording appropriate metadata with a minimum of duplication.

These issues are crucial to creators of digital thematic research collections because 1) the primary resources are held by the archives and libraries, and many of the repositories are digitizing materials and making them available to scholars, even though the digital representations frequently are not as richly represented as those which the scholars will ultimately create; 2) archives and libraries will ultimately be responsible for the collections that are created; and 3) if publishers are to have a role in the publication of digital thematic research collections, standards will be essential for portability, and these standards will by and large come out of the library world.

That said, thematic collections are generally far more complex and dependent on intricate metadata interrelations than digital objects created for most digital libraries. This complexity is due in part to the interpretive aspects of projects in which scholars are deeply engaged with the materials. For example, the poetry manuscripts section of the *Whitman Archive* incorporates images, transcriptions with scholarly annotations, copyright and administrative information, and finding aids. All of these materials are encoded, presented through XSLT stylesheets, and made available via various search options. The Whitman Archive uses an array of data and metadata standards: TIFF, TEI, EAD, and MODS. Because of its complexity, the *Whitman Archive* is a good case study for investigating the use and adequacy of Metadata Encoding and Transcription Standard (METS) as a bundling tool for coordination of the standards found in digital thematic research collections.

In 2005, the Institute of Museum and Library Services awarded a two-year grant to the University of Nebraska-Lincoln to explore interoperability of metadata for thematic research collections and to demonstrate the power of METS as an ingestion package. The current project builds on an earlier grant project to create an integrated guide to dispersed manuscripts. One hypothesis of the current project is that by standardizing the way metadata is encoded, creators of digital thematic research collections can make their work more sustainable and ingestible; in other words, digital scholarship will be collected in a form that will allow libraries and other entities to incorporate into their holdings digital thematic research collections created at diverse locations by a variety of scholars. As the research team's work has progressed, the group has naturally experienced evolution in its thinking and approaches to metadata. Members of the project team will report on project processes and findings.

Organization:

Katherine L. Walter, Co-Director of the Center for Digital Research in the Humanities (CDRH) and co-project director for the interoperability of metadata project, will provide a brief introduction to the topic, including a description of

thematic research collections such as the *Whitman Archive*, the varying types of metadata that may be involved, the impetus to study integration of overlapping standards for such collections, and the project goals originally set by the research team. Walter will introduce speakers, moderate the panel, and invite questions from the audience at the end.

Brett Barney, Senior Associate Editor and Project Manager of the *Walt Whitman Archive*, will describe the team's approach to exploring what constitutes redundancy in metadata and its exploration of questions such as what is really redundancy and what is not? When is it desireable and when is it not? Does redundancy stymie ingestion or aid it? His report will pose conclusions regarding the value of redundancy and its drawbacks in the ingestion of thematic research collections into open-source library catalogs. Barney will also describe the thinking of the team in selecting different kinds of representative digital objects in the Whitman Archive for the ingestion demonstration project.

Julia Flanders, Women Writers' Project, Brown University, will describe issues raised for scholars in contributing all or part of a thematic research collection into library catalogs. Questions explored will include what responsibility does an acquiring library have to preserve or honor a scholar's intentions in creating and shaping such a collection? Is a library in fact ingesting a collection or simply objects in a collection? This will include thoughts on preserving contextual aspects of thematic research collections, navigational features, and interface. As a member of a standards board, she will speak about the value of this kind of research for standards bodies.

Terence Catapano, Special Collections Analyst/Librarian, Libraries Digital Program, Columbia University, will describe the importance and adequacy of METS as a publication and ingestion package, or in OAIS terms, how a Submission Information Package (SIP) is derived for ingestion into a Digital Library, and the development of a METS Profile for Thematic Research Collections. Catapano will briefly describe work of digital library projects that helped inform this study, such as the DLF MODS Implementation Guidelines for Cultural Heritage Materials, Descriptive Metadata Guidelines for RLG Cultural Materials, DLF Aquifer, and others.

Daniel Pitti, Associate Director of the Institute for Advanced Technology in the Humanities will discuss whether metadata should exist in different forms at different stages; and describe ingestion of the *Whitman Archive* into FEDORA, what this means for other thematic research collections, and reflect on the thoughts the FEDORA group at the University of Virginia has with respect to preservation of such collections.

Three Play Effects: Eliza, Tale-Spin, and SimCity

Noah Wardrip-Fruin *(nwf@ucsd@edu)*
University of California San Diego

In the mid-1960s Joseph Weizenbaum created a stunning piece of software. Years before HAL 9000's screen debut in *2001: A Space Odyssey*, this software, *Eliza*, made it possible to have a conversation with a computer. *Eliza's* most famous script, *Doctor*, caused the software to parody the conversational patterns of non-directive therapists during an initial visit. While *Eliza/Doctor* can seem quite smart at first blush, each script for *Eliza* is actually just a set of linguistic tricks. Most of these tricks use keyword-driven "decomposition rules" to take the user's last statement, divide it into pieces, and selectively reuse portions to rephrase it as a question.

But when we interact with a piece of software we don't necessarily get a clear picture of how it actually operates internally. And many users of *Eliza/Doctor* initially developed very mistaken ideas about its internals. Weizenbaum (1976) discusses users who assumed that, since the surface appearance of an interaction with the program could resemble something like a coherent dialogue, internally the software must be very complex. Some at first thought it must be something close to the fictional HAL: a computer program intelligent enough to understand and produce arbitrary human language. This happened so often, and was so striking, that computer science circles developed a specific term for this kind of misunderstanding: "the *Eliza* effect."

This paper is a brief look at the *Eliza* effect, and at two previously-unnamed effects that can arise in the relationship between the surface appearance of a digital system and its internal operations. More specifically, this paper looks where others haven't when exploring versions of this relationship: the area of play.

While the initial experience of *Eliza/Doctor* can create the surface impression of an incredibly complex internal system, sustained interaction with the system, the verbal back-and-forth, invites play ... and linguistic play with *Eliza/Doctor* quickly begins to destroy the illusion. In other words, precisely the open-ended textual interaction that helped foster the illusion of internal complexity and intelligence enables play that draws attention to the system's rote simplicity, its distance from human interaction.

On the other hand, a sort of inverse of the *Eliza* effect can be seen with James Meehan's 1976 *Tale-Spin*, the first major story generation program. *Tale-Spin* generates stories from rules for character behavior and a set of facts about the virtual world. When generating stories in interaction with an audience it asks questions to fill in details about locations, objects, relationships, and so on. In addition, internal *Tale-Spin* mechanisms draw "inferences" from the facts. For example, if it is asserted that a character is thirsty, then the inference mechanisms result in the character knowing she is thirsty, forming the goal of not being thirsty, forming a plan for reaching her goal, etc.

Further, *Tale-Spin* characters can use its inference mechanisms to "speculate" about the results of different courses of action. Meehan's The Metanovel (1976) describes a story involving such speculation, in which a hungry Arthur Bear asks George Bird to tell him the location of some honey. We learn that George believes that Arthur trusts him, and that Arthur will believe whatever he says. So George begins to use the *Tale-Spin* inference mechanisms to "imagine" other possible worlds in which Arthur believes there is honey somewhere. George draws four inferences from this, and then he follows the inferences from each of those inferences, but he doesn't find what he's after. In none of the possible worlds about which he's speculated is he any happier or less happy than he is now. Seeing no advantage in the situation for himself, he decides, based on his fundamental personality, to answer. Specifically, he decides to lie.

This is a relatively complex piece of psychological action, and certainly tells us something about George as a character. But the surface output of a *Tale-Spin* story never contains any information about this kind of action. No matter how creatively one plays with *Tale-Spin*, such hidden action cannot be deduced from its surface. This is probably why, though *Tale-Spin* is seen as a landmark in computer science circles, it is often treated with near-ridicule in literary circles. Janet Murray, Espen Aarseth, Jay David Bolter, and other critics have failed to see what makes *Tale-Spin* interesting, focusing instead on what its output looks like on the surface. Or, to put it another way, *Tale-Spin* fails to display its interesting internal processes in a manner that makes them visible to even the most careful of critics.

This situation is far from uncommon in digital media, perhaps particularly in the digital arts, where fascinating processes — drawing on inspirations ranging from John Cage to the cutting edge of computer science — are often encased in an opaque surface. In fact, this effect is at least as common as the *Eliza* effect, though I know of no term that describes it. Given this, I propose "the *Tale-Spin* effect" as a term for works that appear, on their surface, significantly less complex than they are internally.

An effect quite different from both of these can be seen in the case of Will Wright's 1989 game *SimCity*. The seed for this project was planted as Wright created a landscape editor for authoring his first game, an attack helicopter simulation. Working with the editor, he realized he was having more fun making virtual spaces than blowing them up. From this the idea for Wright's genre-defining *SimCity* was born.

SimCity, of course, unlike a terrain editor, doesn't simply wait for a user to do something. Time begins passing the moment a new city is founded. A status bar tells the player what's needed next — starting with basic needs like a residential zone and a power plant and, if play succeeds for any period, ramping up to railroads, police stations, stadiums, and so on. As cities grow, areas respond differently. Some may be bustling while others empty out, or never attract interest. *SimCity* provides different map views that can help diagnose problems with abandoned areas. Players can try changing existing areas of the city (e.g., building additional roads) or create new areas with different characteristics. Observation and comparison offer insights, while answers are found by trying different approaches and considering the results.

In other words, the process of play with *SimCity* is one of learning to understand the system's operations. Conversely, as Wright explains, the challenge of game design is to create a surface experience that will make it possible for audiences to build up an appropriate model of the system internals.

Here, again, we lack a term for an experience. I propose "the *SimCity* effect" for this important phenomenon: a system that, through play, brings the player to an accurate understanding of the system's internal operations. Of course, the *SimCity* effect is named for cases where the system is complex, but the phenomenon can be observed generally. *Pong* works as well as it does because it effectively communicates at the surface level its quite simple internal operations.

What is exciting about the *SimCity* effect, and about Wright's work generally, is that it helps us get at the new possibilities opened by working with computational media. *Pong* is very similar to games we play without computers, but *SimCity* is a more complex system than even the most die-hard Avalon Hill fan would want to play as a tabletop game. This ability to work with computational processes, to create complex computational systems, is the opportunity that digital media affords — and the *SimCity* effect points the way toward creating experiences of this sort that succeed for audiences.

Bibliography

Aarseth, Espen J. *Cybertext: Perspectives on Ergodic Literature.* Baltimore: Johns Hopkins University Press, 1997.

Bolter, Jay David. *Writing Space: The Computer, Hypertext, and the History of Writing.* New Jersey: Lawrence Erlbaum Associates, Inc, 1991.

Meehan, James R. PhD thesis. Yale University, 1976.

Murray, Janet H. *Hamlet on the Holodeck.* New York: The Free Press, 1997.

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation.* New York: W.H. Freeman, 1976.

The Master Builders: LAIRAH Research on Good Practice in the Construction of Digital Humanities Projects

Claire Warwick *(c.warwick@ucl.ac.uk)*

School of Library, Archive and Information Studies
University College London

Melissa Terras *(m.terras@ucl.ac.uk)*

School of Library, Archive and Information Studies
University College London

Paul Huntington *(p.huntington@ucl.ac.uk)*

School of Library, Archive and Information Studies
University College London

Nikoleta Pappa *(n.pappa@ucl.ac.uk)*

School of Library, Archive and Information Studies
University College London

Isabel Galina *(i.russell@ucl.ac.uk)*

School of Library, Archive and Information Studies
University College London

Abstract:

This paper describes the results of research carried out during the LAIRAH (Log analysis of Internet Resources in the Arts and Humanities) project (<http://www.ucl.ac.uk/slais/circah/lairah/>) which is based at UCL's School of Library Archive and Information Studies. It was a fifteen month study (reporting in October 2006) to discover what influences the long-term sustainability and use of digital resources in the humanities through the analysis and evaluation of real-time use.

At Digital Humanities 2006 we reported on the early stages of the project, in which we carried out deep log analysis of the AHDS and Humbul portals to determine the level of use of digital resources. (Warwick et al. 2006) This proposal will discuss the results of the final phase of the research in which we examined digital resources from the point of view of those who designed and built them. We aimed to discover whether

there were common characteristics and elements of good practice linking resources that are well- used.

Numerous studies have been carried out into the information needs and information seeking practices of humanities scholars (Barrett, (2005) Talja and Maula (2003), Herman (2001) and British Academy, (2005)). However, our research is original because it surveys the practices of those who produce digital humanities resources. We also based the selection of our projects on deep log analysis: a quantitative technique which has not previously been applied to digital humanities resources to ascertain real usage levels of online digital resources.

Method:

We selected a sample of twenty one publicly funded projects with varying levels of use, covering different subject disciplines, to be studied in greater depth. We classified projects as well-used if the server log data from the Arts and Humanities Data Service (AHDS) and Humbul portals showed that they had been repeatedly and frequently accessed by a variety of users. We also mounted a questionnaire on these sites and asked which digital resources respondents found most useful. Although most nominated information resources, such as libraries, archives and reference collections for example the eDNB, three UK publicly funded research resources were mentioned, and thus we added them to the study. We also asked representatives of each AHDS centre to specify which resources in their collections they believed were most used. In the case of Sheffield University the logs showed that a large number of digital projects accessed were based at the Humanities Research Institute. We therefore conducted interviews about the HRI and its role in fostering the creation of digital humanities resources.

The selected projects were studied in detail, including any documentation and reports that could be found on the project's website. We also interviewed a representative of the project, either the principal investigator or a research assistant.

Results:

Institutional context:

The majority of projects that we interviewed had been well supported in technical terms, and this had undoubtedly aided the success of the project, especially where it was associated with a centre of digital humanities excellence such as the Centre for Computing in the Humanities at Kings College London or the HRI at Sheffield. Critical mass aided the spread of good practice in the construction and use of digital resources in the humanities. Where a university valued such activities highly they tended to proliferate. More junior members of staff were inspired to undertake digital humanities research by the success of senior colleagues and early adopters respected for their traditional and digital research. Unfortunately such critical mass is relatively rare in UK universities and some PIs reported that their digital resource was not understood or valued by their departments, and thus their success had not lead to further digital research.

Staffing:

PIs also stressed how vital it had been to recruit the ideal RAs. These were however relatively difficult to find, as they had to have both disciplinary research expertise and good knowledge of digital techniques. Most RAs therefore required training, which many PIs often found lacking or of poor quality. A further frustration was the difficulty of finding funding to continue research, this meant that an expert RA might leave, necessitating further training of a new employee if the project was granted future funding.

Dissemination:

The strongest correlation between well-used projects and a specific activity was in the area of dissemination. In all the projects studied, staff had made determined efforts to disseminate information as widely as possible. This was a new challenge for many humanities academics, who were more used to writing books, marketed by their publishers. This might include giving papers at seminars and conferences both within the subject community and the digital humanities domain; sending out printed material; running workshops, and in the most unusual instance, the production of a tea-towel!

User contact:

Very few projects maintained contact with their users or undertook any organised user testing, and many did not have a clear idea how popular the resource was or what users were doing with it. However, one of the few projects that had been obliged to undertake user surveys by its funders was very well-used, and its PI had been delighted at the unexpected amount and range of its use. Another project came to the belated realisation that if it had consulted users the process of designing the resource would have been simpler and less demanding.

Documentation:

Few of the projects kept organised documentation, with the exception of those in archaeology, linguistics and archival studies, where such a practice is the norm in all research. Most projects had kept only fragmentary, internal documents, many of which would not be comprehensible to someone from

outside. Documentation could also be difficult to access, with only a small minority of projects making this information available from its website. This is an important omission since documentation aids reuse of resources, and also provides vital contextual information amount its contents and the rationale for its construction that users need to reassure them about the quality of the resource for academic research.

Sustainability:

Another area of concern was the issue of sustainability. Although the resources were offered for deposit with the AHDS, few PIs were aware that to remain usable, both the web interface and the contents of the resource would require regular updating and maintenance, since users tend to distrust a web page that looks outdated. Yet in most cases no resources were available to perform such maintenance, and we learnt of one ten year old resource whose functionality had already been significantly degraded as a result.

Conclusion and recommendations

Well-used projects do therefore share common features that predispose them to success. The effect of institutional and disciplinary culture in the construction of digital humanities projects was significant. We found that critical mass was vital, as was prestige within a university or the acceptance of digital methods in a subject. The importance of good project staff and the availability of technical support also proved vital. If a project as to be well-used it was also essential that information about it should be disseminated as widely as possible.

Even amongst well-used projects, however we found areas that might be improved, these included organised user testing, the provision of and easy access to documentation and the lack of updating and maintenance of many resources.

Recommendations:

Documentation:

- Projects should keep documentation and make it available from the project web site, making clear the extent, provenance and selection methods of materials for the resource.
- Funding bodies might consider making documentation a compulsory deliverable of a funded project.
- Discussions could be held between relevant stakeholders and the funding bodies, with the aim of producing an agreed documentation template. This should specify what should be documented and to what level of detail.

Users:

- Projects should have a clear idea of whom the expected users might be; consult them as soon as possible and maintain contact through the project via a dedicated email list , website feedback or other appropriate method
- They should carry out formal user surveys, software and interface tests and integrate the results into project design.
- Applicants for funding should show that they have consulted documentation of other relevant projects and discuss what they have learnt from it in their case for support. The results of such contact could then be included in the final report as a condition of satisfactory progress.

Management:

- Projects should have access to good technical support, ideally from a centre of excellence in digital humanities.
- Projects should recruit staff who have both subject expertise and knowledge of digital humanities techniques, then train them in other specialist techniques as necessary.
- Funding bodies might consider requiring universities to offer more training for graduate students and RAs in digital humanities techniques.

Sustainability:

- Ideally projects should maintain and actively update the interface, content and functionality of the resource, and not simply archive it with a data archive such as the AHDS. However this is dependent on a funding model which makes this possible.

Dissemination:

- Disseminate information about itself widely, both within its own subject domain and in digital humanities.
- Information should be disseminated widely about the reasons for user testing and its benefits, for example via AHRC/AHDS workshops. Projects should be encouraged to collaborate with experts on user behaviour.

Acknowledgements:

This project was funded by the Arts and Humanities Research Council ICT Strategy Scheme. We would also like to thank all of out interviewees for agreeing to talk to us.

Bibliography

Barrett, A. "The Information Seeking Habits of Graduate Student Researchers in the Humanities." *The Journal of Academic Librarianship* 31.4 (2005): 324-331.

British Academy. *E-resources for Research in the Humanities and Social Sciences - A British Academy Policy Review section 3.5* . 2005. <http://www.britac.ac.uk/reports/eresources/report/sect3.html#part5>

Herman, E. "End-users in Academia: Meeting the Information Needs of University Researchers in an Electronic Age Part 2 Innovative Information-accessing Opportunities and the Researcher: User Acceptance of IT-based Information Resources in Academia." *Aslib Proceedings.* 2001. 431-457.

Talja, S., and H. Maula. "Reasons for the Use and Non-use of Electronic Journals and Databases - A Domain Analytic Study in Four Scholarly Disciplines." *Journal of Documentation* 59.6 (2003): 673-691.

Warwick, C., M. Terras, P. Hungtington, and N. Pappa. "If You Build It Will They Come? The LAIRAH Survey of Digital Resources in the Arts and Humanities." Paper presented at Digital Humanities 2006, Paris Sorbonne, 5-9 July 2006. 2006.

The KWIC-step: A Dance for 2 or More

Susan L. Wiesner *(dap2sw@surrey.ac.uk)*
University of Surrey

This cross-discipline paper considers the use of corpus linguistics as it applies to a study of dance writing.

Argument/Premise

Linguists and Computational Linguistic Engineers who have studied language for special purposes (LSP) generally agree that the lexicon used by those in a specialist discipline offers insight into the concepts and ideologies of that discipline. As Dance is a specialist discipline, it should then follow that the lexicon of dance writing can demonstrate aesthetic concepts and theoretical approaches and beliefs through generalized language patterns detected using KWIC methods. This research supports that contention in that the empirical approach (specifically key words in context), when used against a corpus of 1.4 million words from written dance texts, does offer a means to develop an ontology of the discipline of dance. One ontological example supported by the empirical data is that of both hierarchical and rhizomatic structures of relationships between participants in the dance event. Through this example, this paper will demonstrate both the process and analytical product of using corpus linguistics methods and models to develop an ontology of Dance.

Analytical Method

Due to a dearth of research into dance writing, especially in the area of corpus linguistics, there were no tested, dance-specific models on which to base this study. Therefore, this research contains a variety of methodologies and approaches in the disciplines of corpus linguistics and dance analysis. The corpus methodologies are based on the work of Sinclair (2003), Ahmad (2002), and Biber (1998). Linguistic analyses based on KWIC (key words in context) such as word frequencies, collocations, concordances, and POS tagging were performed on the data generated by computing tools (e.g. System Quirk, Unitex, CLAWS). Additional methods proposed by Traboulsi et al (2003) for identifying a local grammar were used against samples of critical texts (i.e. dance performance reviews). Finally, through a two-pronged approach, the data

was analysed using the analytical modes of description, interpretation, and evaluation proposed for dance by Adshead et al (1988).

After establishing the written language of Dance as a specialist language (or language for special purposes) through the frequency of open class words in the top 100 most frequently used words, weirdness factors, and Z-scores, the data was analysed in a variety of ways. For example, for the purposes of this particular study into ontological relationships between participants, general concepts were identified by focusing on those open class words (OCWs) within the top 100 most frequently used words in a general dance corpus (the Surrey Dance Corpus, or SDC). Finding that the OCWs included several conceptual possibilities which appeared at times to cross conceptual boundaries, it became apparent that a method for categorisation was required to provide distinctions between the concepts. Therefore, the analytical model introduced by Adshead et al (1988) was used for the categorisation. This model calls for a deconstruction of a dance work via description, interpretation, and evaluation based upon identifiable characteristics of the individual work. The identified distinctions include not only the movement and production components, but also concepts of form, concepts through which we view a dance, and concepts particular to a dance. After determining the conceptual categories, the analysis went beyond the top 100 most frequently used words, and considered the relative frequencies of the concept words in the subsequent groups of most frequently used words (top 200, 300, etc.). Word counts also were performed on personal pronouns and conceptual metaphor (using Charteris-Black's method). Finally, collocations and concordances using the concept words as target words were generated and analysed.

Each step was repeated using additional corpora: a group of sub-corpora defined by content of the texts and/or writers' particular activity within the discipline (e.g. critic, scholar, choreographer); and a corpus focused on one choreographer (used as a case study throughout the research).

Conclusion

Through this research into dance writing, Dance can be shown to be a LSP with a lexicon specific to the discipline. So, too, have several unexpected patterns been detected that reflect various concepts and ideologies, one of which is that of the relationships between and among the participants in the dance event (e.g. dancer, choreographer, audience, critic, company). The relational connections are further supported by a general belief in the collective, as shown through word counts and concordances. Additional methods used in dance analysis provide a structure for studying and further clarifying the empirically generated patterns. The choreographic and analytic concepts in Dance as they are perceived and expressed through written dance texts demonstrate not only the importance of these concepts to specialists in the discipline of Dance, but also the benefit of using the corpus linguistics approach to studying dance writing.

Bibliography

Adshead, J., V. A. Bringinshaw, P. Hodgens, and M. Huxley. Ed. J. Adshead. *Dance Analysis: Theory and Practice.* London: Dance Books, 1988.

Ahmad, K. "The Role of Specialist Terminology Management in Artificial Intelligence and Knowledge Acquisition." *Applications Oriented Terminology Management.* Ed. S. E. Wright and G. Budin. Amsterdam: John Benjamins, 2001. 809-844.

Biber, D, S. Contrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press, 1998.

Charteris-Black, J. . *Corpus Approaches to Critical Metaphor Analysis.* New York: Palgrave MacMillan, 2004.

Sinclair, J. *Reading Concordances.* London: Pearson Education Limited, 2003.

Trabousli, H., D. Cheng., and K. Ahmad. 2003.

The Abbey Inside the Machine: The MonArch Project

Clifford Edward Wulfman
(Clifford_Wulfman@brown.edu)
Brown University
Elli Mylonas *(elli_mylonas@brown.edu)*
Brown University
Anne Loyer
Wesleyan University
Sheila Bonde
Brown University
Clark Maines
Wesleyan University

Description of Project

MonArch (The Monastic Archaeology Project) is a collaborative project based on the excavations at the Abbey of Saint-Jean-des-Vignes in Soissons, France under the joint directorship of Sheila Bonde (Brown University) and Clark Maines (Wesleyan University). The monastery at St.-Jean-des-Vignes remains with us today in two media: in the physical remains of the abbey and in the textual remains left behind by its inhabitants. The most significant of the latter are the community's Customary, which describes the roles and activities of the abbey's inhabitants over the course of the religious year, and the Abbey's Obituary, which lists, by month and day of death, the members of the monastic community and laics who had provided service or gifts to the monastery, with a description of their donations.

Two websites have been developed as part of the MonArch Project. The first was implemented at Wesleyan (<http://www.wesleyan.edu/monarch/>), and provides thorough documentation of the excavations, annual field reports, and inventories of finds. The second, which is the focus of this paper, is a research effort in its own right, one that explores the ways in which encoded representations of archaeological data may be used enhance their use. The often complex interrelationships among textual, architectural, and archaeological evidence are difficult to represent and explicate in traditional formats; the work described here attempts to use digital representations to create a new form of scholarly expression.

Research Questions

MonArch has always focused beyond a simple description of concrete artifacts to a thoroughgoing investigation and elaboration of a concept: from the outset, the project was conceived as a multi-disciplinary endeavor that took *monasticism* rather than the monastery as its object of study.[1] Therefore, to the extent that its digital manifestation is to be a direct reflection of the project's aims, it too must take as its focus an abstraction and not a collection of materials.

Such a focus makes designing the digital reflection more challenging. It cannot be a simple linked, digital collection, because such an organization would put undue emphasis on the monastery and its artifacts. And because the site needs to be built around concepts and questions — about monasticism and the archaeology of monasticism — the digital project cannot be conceived or designed without considerable knowledge of those concepts or an understanding of medieval social, economic, architectural, and religious history and the role of archaeology in elucidating it all. It must, in other words, express the *archaeological definition of monasticism* Bonde and Maines are deriving from their work, a definition that is based on the following factors:

- The monastic complex as the physical expression of spiritual, social, and economic motives;
- The functions of the abbey's structures;
- The communitas who used these structures and the routine of their lives;
- The economy of the surrounding region — the farms, mills, priories, parishes, and other holdings that constituted the abbey's domain;
- The place of the abbey (defined as both buildings and community) in the local, regional, and inter-regional networks of power and influence;
- The way these phenomena changed over time.

The relationship of the texts — both primary and secondary — to the archaeological data is complex, because the archaeological material supports and illustrates the texts at the same time that the texts contextualize the archaeology and architecture. These two kinds of remains — the textual and the physical — become two inter-referential and mutually reinforcing centers from which the concepts of monasticism emerge. The research focus of the current digital project is on designing an infrastructure and and an interactive interface that simultaneously represent the authors' arguments and allow users of the site to form their own interpretations.

Considering Use

We argue that it is most fruitful to conceive the digital manifestation of the MonArch project as a *reading*. Insofar as its goal is to provide a definition, the resource is fundamentally rhetorical: it has an argument, and it supports its claims with evidence drawn from a variety of evidential sources. That argument, generally stated, entails an archaeological definition of monasticism, whose refinement is the ongoing and iterative task of the MonArch project. The task of the digital resource, therefore, must be unusually subtle: it must express this archaeological definition through an intricate blend of textual and material evidence, and it must provide a way for readers to interact with and question the argument and its claims.

In fact, insofar as the digital resource is the manifestation of this argument, the mutually referential presence of both encoded texts and artefactual simulacra constitutes a claim in itself, a claim that a 'definition' of monasticism can be manufactured through a reading of monastic text, monastic space, and monastic time, and perhaps only in this way. That is to say, the boundaries established by the site (and here the web site and the excavation site perhaps begin themselves to lose their boundaries) determine how monasticism is to be understood.

Research Goals

Our research, then, is focused on establishing a layered set of *models* that enable researchers to articulate their understanding of monasticism and allow scholars (students, readers, users) to interact with that understanding. Underlying the whole project is a fundamental data model that represents the characteristics of the texts, spaces, buildings and artifacts that form the object of study. Overlaying the data model is another model, that of the relationships that embody the intellectual connections that the researchers have made as they work over their materials and which embody their definition of monasticism. These connections are the evidence for their claims. At the topmost layer is a model of user interaction, one consisting of visual juxtapositions that illustrate the relationships among the historical evidence and that enable further questioning.

The problem of structuring information in order to enhance argument is part of a larger problem. One of the major powers promised by digital resources is the instantiation of the kind of textuality envisioned by postmodern theorists like Roland Barthes and Jacques Derrida, among others: the simultaneous availability of vast amounts of information in a form that makes the interconnections, both explicit and implicit, traversable. Yet benightedness is a clear danger: it is all too easy to become lost: in 'hyperspace,' in the library, in the labyrinth, or in the wood of error. As digital resources grow in size and complexity, the need for prospects — lookouts, overviews of the textscape — becomes ever greater. Digital resources must therefore become responsive: when a reader examines an argument or claim in a digital publication, the resource should respond to her, helpfully putting on the virtual desk before her materials that are relevant to her evolving inquiry. Thus, in loftiest terms, our goal must be to assist, supplement, and augment a human agent 's investigation of accumulated cultural knowledge, a goal congruent with the aims of research over the past half-century in fields from information retrieval to artificial intelligence.

---

1. Sheila Bonde and Clark Maines, *Saint-Jean-des-Vignes in Soissons: Approaches to its Architecture, Archaeology and History*, Biblioteca Victorina vol. XV (Turnhout: Brepols Press, 2003).

Bibliography

McCarty, Willard. *Humanities Computing.* Basingstoke: Palgrave McMillan, 2005.

Bernstein, Mark, J. David Bolter, Michael Joyce, and Elli Mylonas. "Architectures for Volatile Hypertexts." *Hypertext '91 Proceedings.* San Antonio, TX: Association for Computing Machinery, 1991.

Bonde, Sheila, and Clark Maines. "The Archaeology of Monasticism, Ten Years of Work at the Augustinian Abbey of Saint-Jean-des-Vignes, Soissons." *Medieval Europe, 1992 (Pre-printed Papers of the Conference on Medieval Archaeology in Europe, York, 21-24.IX.1992).* York, 1992. 83-88.

Bonde, Sheila, and Clark Maines. *Saint-Jean-des-Vignes: Approaches to its Architecture, Archaeology and History.* Biblioteca Victorina vol. XV. Turnhout: Brepols Press, 2003.

Wulfman, Clifford, Julia Flanders, and Elli Mylonas. "The Rhetoric of Performative Markup." *Digital Humanities 2006 Conference Abstracts.* Paris: CATI, Université Paris-Sorbonne, 2006. 248-251.

Monarch. Monastic Archaeology Project . <http://www.wesleyan.edu/monarch/>

The Abbey of St.-Jean-des-Vignes . <http://dev.stg.brown.edu/projects/Monarch/>

Unsworth, John. "What is Humanities Computing and What Is Not?" <http://www.computerphilologie.uni-muenchen.de/jg02/unsworth.html>

An Evaluation of Text Classification Methods for Literary Study

Bei Yu *(beiyu@uiuc.edu)*
University of Illinois at Urbana-Champaign
John Unsworth *(unsworth@uiuc.edu)*
University of Illinois at Urbana-Champaign

A survey study [1] shows that text classification is a typical scholarly activity in literary study, and automatic text classification methods can be used in three scenarios. The first is information organization - a classifier can learn the target category concepts (e.g. news article about trade, acquisition, etc.) from the training documents, and then assign new documents into these predefined categories. The second purpose is knowledge discovery - a successful classifier can provide insights to understand a target concept by revealing the correlations between the features and the concept. The third purpose is example-based retrieval - a classifier might be able to learn a concept from a small number of training documents with the help of semi- supervised learning or active learning methods, and then retrieve more documents similar to the training examples from a large collection.

Text classification techniques have been well developed in the past twenty years. With the availability of many text classification methods, empirical evaluation is important to provide guidance for method selection in applications. Because of the sub jectivity in the class concept definition, analytical evaluation of text classifiers is difficult. Therefore empirical experiments became the common text classification evaluation methods.[2] The major text classification methods have been evaluated on topic classification tasks using some benchmark data sets, such as the Reuters-21578 news collection and the Usenet newsgroup collection. Some topic classification evaluation results have been widely accepted. For example, SVM is currently the best text classifier[3]; no feature selection improves SVM performance [4]; SVM feature selection is better than Odds Ratio for naive Bayes, etc.[5] There are mixed conclusions regarding some document preprocessing techniques, such as stemming and stop word removal.

However, these evaluation data sets were limited to news and web documents; the evaluation tasks were limited to topic classification for information organization purpose. The target concepts in literary text classification range from topic to style, genre, emotion, and more. These different types of target concepts can also be called document properties. Previous study [6] showed that document properties interact with clustering methods. Will the various document properties in literary text classification tasks affect the classification methods? Are the previous evaluation results still valid for literary text classification?

This paper describes an empirical evaluation of text classification methods for literary study. We choose a new kind of data - the literary documents - to evaluate classification methods. Because no benchmark data is available in the literary domain, we select two literary text classification problems - the eroticism classification in Dickinson's poems and the sentimentalism classification in early American novels - as two cases for this study. Both problems focus on identifying certain kinds of emotion - a document property other than topic.

We also choose two popular text classification algorithms - naive Bayes and Support Vector Machines (SVM), and three feature engineering options - feature merging (stemming), stopword removal and statistical feature selection (Odds Ratio and SVM) - as the sub jects of evaluation. We aim to examine the effects of the chosen classifiers and feature engineering options on the two emotion classification problems, and the interaction between the classifiers and the feature engineering options. As a special case of feature merging, we also examine the impact of Dickinson's unconventional capitalizations on classification performance. We choose bag-of-words (BOW) model for document representation.

We seek empirical answers to the following research questions:

1. 1. Is SVM a better classifier than naive Bayes regarding classification accuracy, new literary knowledge discovery and potential for example-based retrieval?
2. 2. Is SVM a better feature selection method than Odds Ratio regarding feature reduction rate and clas- sification accuracy improvement?
3. 3. Does stop word removal affect the classification performance?
4. 4. Does stemming affect the performance of classifiers and feature selection methods?

Our experiment results show that SVM is not a universal winner in literary text classification. After fea- ture reduction naive Bayes achieves high accuracies in both cases while SVM succeeds in the sentimentalism classification only. Figure 1 and 2 show that SVM and naive Bayes select their top features from different frequency ranges. Naive Bayes tends to pick unique words, which are often not frequent. The large number of low frequency words results in the success of naive Bayes in the eroticism classification. These unique words also surprised the Dickinson scholars, who finally found some new erotic indicators from them. SVM tends to pick high frequent

and discriminant words, which are scarce in the Dickinson collection. These words (such as personal pronouns) are within the scholars' expectation and therefore not interesting anymore.

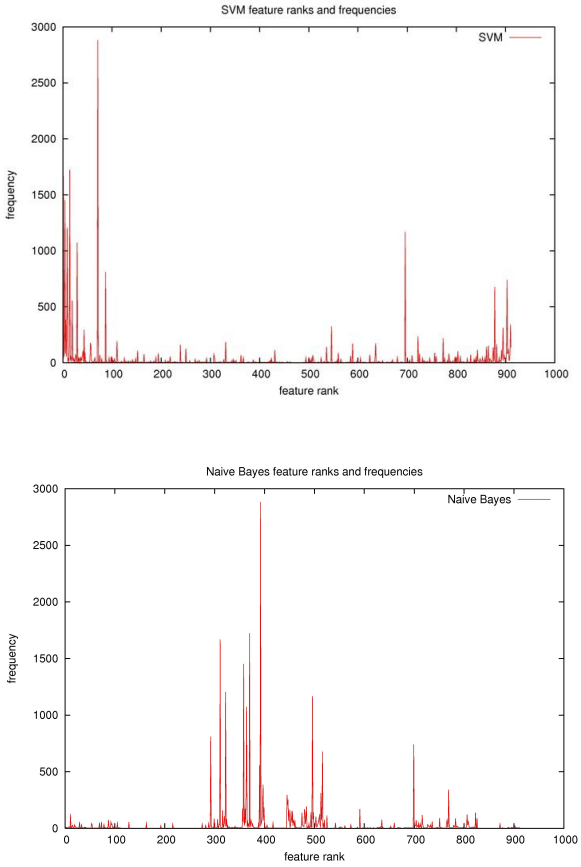


Figure 1: Dickinson feature ranks and frequencies

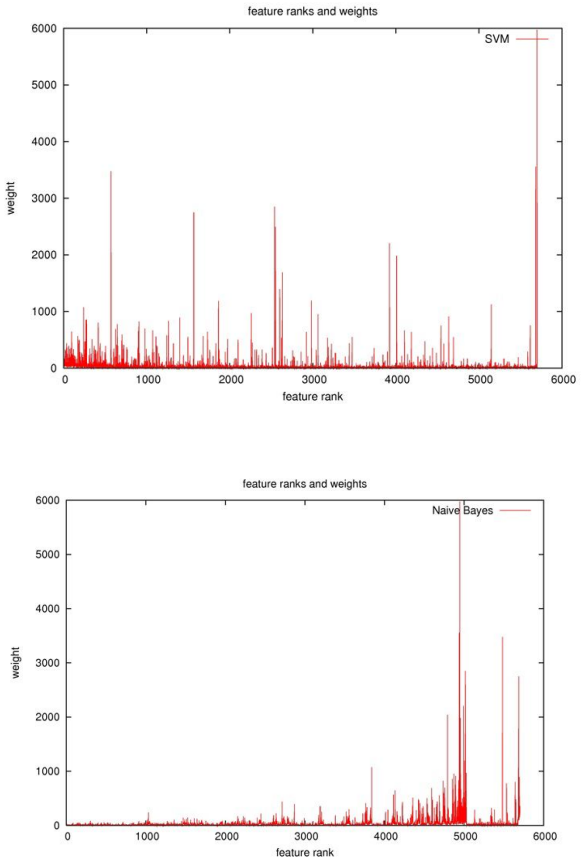


Figure 2: Sentimentality feature ranks and frequencies

Despite the high classification accuracies, the learned naive Bayes eroticism classifier, and also the SVM classifier with low accuracy, are useless for example-based retrieval purpose. Figure 3 shows that the concept of eroticism can not be learned from small number of examples, and the classifiers' prediction confidence drop quickly with the expanding prediction coverage.

Both classifiers achieve high classification accuracies in the sentimentalism classification task, which in- dicate that sentimentalism is a more straightfoward concept than eroticism for bag-of-words representation. The two classifiers still choose different top features but reach comparable performances for the sentimen- talism classification. So for the purpose of feature-category correlation analysis the two methods should be used as complemental to each other rather than one over the other. This time the unique words picked by naive Bayes are so strange that the scholars can not make sense of it. The common but discriminant words picked by SVM are still within the scholars' expectation. The learning curves and confidence curves in figure 4 show that both classifiers yield high potential for example-based retrieval.

The experiment results also show that self feature selection helps both naive Bayes and SVM improve classification accuracies. For SVM the improvement is not as significant as for naive Bayes. Odds Ratio is better than SVM as feature selection method for naive Bayes. However Odds Ratio cannot improve the SVM performance. Without feature selection the stemmed and unstemmed features obtain similarly low classification accuracies in both cases, so did the case merging in the Dickinson case. The micro level analysis finds that the effects of good mergings and bad mergings are neutralized overall. Stemming does not affect both feature selection methods in the eroticism classification case, but we are surprised to find that stemming negatively affects both feature selection methods, especially SVM, in the sentimentalism classification case.

We have found that the stop words obtained from the Brown corpus are also overly common and useless in sentimentalism classification. However, the Brown stop words are mostly uncommon in the Dickinson collection. Personal pronouns - the group of function words usually treated as stop words - turns out to be highly relevant features for eroticism classification.

Our study extends the empirical evaluation of text classification methods to emotion classification tasks in the literary domain. Some conclusions are consistent with what are obtained in previous research, such as Odds Ratio does not improve SVM performance and stop word removal might harm classification. Some conclusions contradict previous results, such as SVM does not beat naive Bayes in both cases. Some findings are new to this area - SVM and naive Bayes select top features in different frequency ranges; stemming might harm feature selection methods. These experiment results provide new insights to the relation between classification methods, feature engineering options and non-topic document properties.

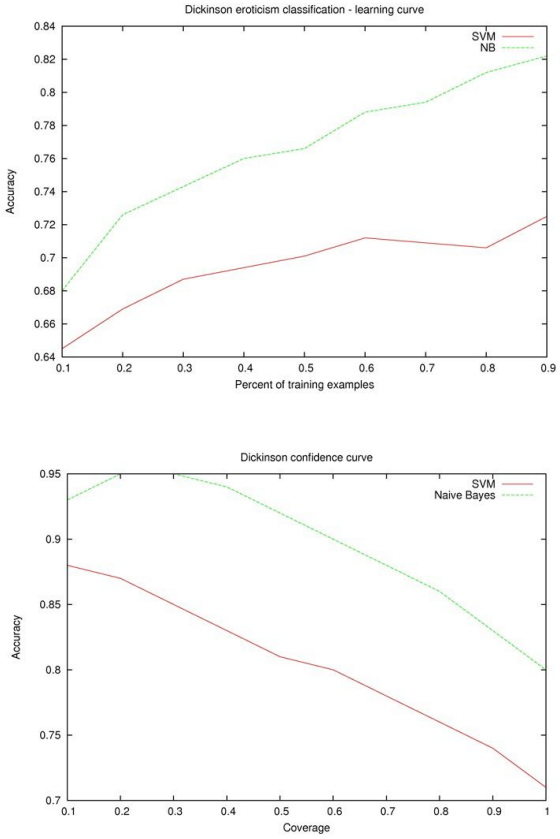


Figure 3: Figure: potential for Dickinson example-based retrieval

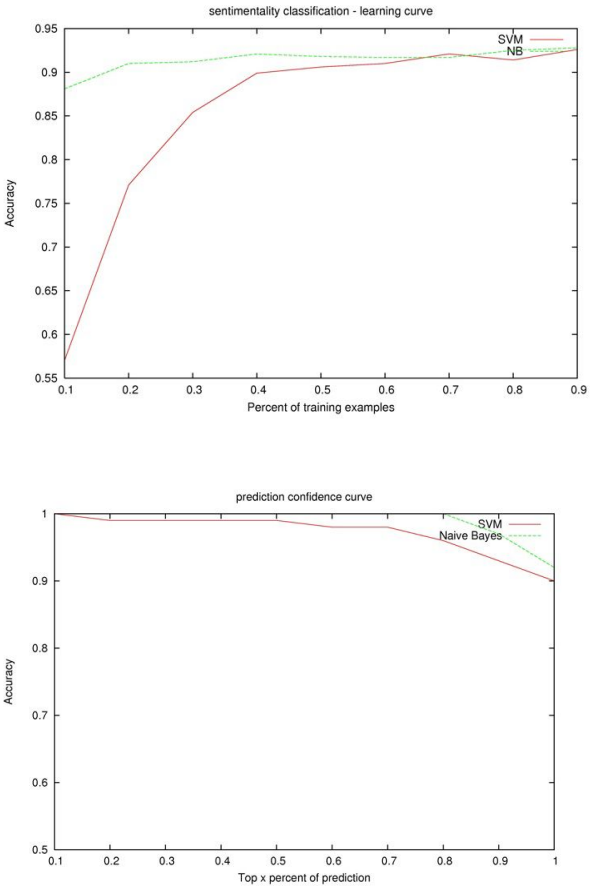


Figure 4: Figure: potential for sentimentalism example-based retrieval

Our experiment results also provide guidance for classification method selection in literary text classifi- cation applications. We suggest that both SVM and naive Bayes be used for feature-category correlation analysis purpose. The number of support vectors in the SVM model indicates the complexity of the target concept. A complex concept is hard to learn from small training set. Feature reduction produces smaller and more generalizable models, but statistical methods are a better choice than the arbitrary feature reduction (like stemming and stop word removal) which are insensitive to particular classification tasks.

1. Bei Yu and John Unsworth, " Toward Discovering Potential Data Mining Applications in Literary Criticism," *Digital Humanities 2006 Conference Abstracts* (Paris: CATI, Université Paris-Sorbonne, 2006): 237-239.
2. Fabrizio Sebastiani, " Machine Learning in Automated Text Categorization," *ACM Computing Surveys* 34.1 (2002): 1-47.
3. Thorsten Joachims, " Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proceedings of ECML-98, 10th European Conference on Machine Learning* 1998, and Yiming Yang and Xin Liu, " A Re-evalution of Text Categorization Methods," *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99),* (Berkley, CA: AMC, 1999): 42–49
4. George Forman, " An Extensive Empirical Study of Feature Selection Metrics for Text Categorization," *Journal of Machine Learning Research* 3 (2003): 1289–1305, and Dunja Mladenic and Marko Grobelnik, " Feature Selection for Unbalanced Class Distribution and Nave Bayes," *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)* (199): 258–267.
5. Dunja Mladenic, Janez Brank, Marko Grobelnik, and Natasa Milic-Frayling, " Feature Selection Using Linear Classifier Weights: Interaction with Classification Models," *ACM SIGIR '04* (2004): 234–241.
6. J. Morato, J. Llorens, G. Genova, and J. A. Moreiro, " Experiments in Discourse Analysis Impact on Information Classification and Retrieval Algorithms," *Information Processing and Management* 39 (2003): 825– 851.

RolandHT and Reconceiving the Notion of Corpus

Vika Zafrin *(vika@wordsend.org)*
Brown University

RolandHT

For the past several years I have been pursuing the study of a semi-fictional character's manifestations in different art forms, throughout Europe and North America, in the last thousand years. Roland is the protagonist of the French epic *Song of Roland* and France's national hero. He first appeared in written record at the end of the eleventh century, and hasn't left us since then – although his character has undergone significant changes. He has easily passed from one culture to another; but Roland is not merely borrowed – he is mercurial. He has been used by writers and artists as a container for the promotion of moral and political exigencies relevant to their circumstances. Thus, in France he is a staunch protector of the Christian faith (*Song of Roland*); in Germany, a religious martyr (Konrad); in Italy, a lover and a comedic character (Boiardo, Ariosto); in England, a mystical traveler (Browning); in the United States, a roadie (Silverstein) and a gunslinger (King).

Nevertheless, the French Roland has easily identifiable personality traits, such as his infamous pride (which cost him and twenty thousand of Charlemagne's best men their lives) and his physical prowess. There are also artifacts, such as his enchanted sword and his sounding horn that can be heard for miles. These features have prompted me to make a case for the existence of a Roland corpus that is held together by recurrent themes and imagery. In my dissertation titled *RolandHT*I show these recurrences by encoding them in XML.[1] This corpus has never been considered as a whole. An examination of themes recurrent in it offers insight into the evolution of art and the pace of intercultural transmission. This insight is relevant to the study of literature at the dawn of Western secular writing; but it also sheds light on our own literary present, in an age of widely accessible electronic writing, when the processes of composition and transmission are changing again.

The Roland corpus consists of works in such diverse media as literature, live and puppet theater, opera, sculpture and stone carvings, painting, film, contemporary music, and comics. This presents a problem in my use of the word *corpus*, which conforms to its theoretical definition but not to any of its prior use in practice. I propose that the electronic medium, and semantic encoding in particular, allow for the study of corpora bound not by format or authorship, but by semantic content.

RolandHT is a collection of excerpts from works that feature Roland, encoded for recurring characters, imagery and themes. Some annotations are present. Works span most of Europe and the United States, and date from between the late 1090s to 1995. Most of the corpus objects are text-based – poems, prose, drama, operatic lyrics. One particularly interesting corpus object – Greg Roach's computer game *The Madness of Roland* – is a blend of text and images. Other image-based corpus objects include comic-book panels and photographs of medieval stonework found in France and Croatia. Included are also excerpts from the 1999 film *Beowulf*, in QuickTime (.mov) format. Other objects will be added as time and copyright restrictions allow.

Technical Details

The project is processed for web presentation using XSLT, JavaScript and CSS. The website is a technical collaboration with Ethan Fremen.

Our approach is standards- and client-based. All features except bookmarking can be used with static files accessed locally. This permits *RolandHT* to be archived on a CD, DVD or other storage device; access to a networked server is not necessary to view the work. This is an advantage for the purposes of submitting a doctoral thesis, but being entirely client-side, our method would not scale well: a substantially larger corpus will require an XML database in order to remain efficient. As it stands, the roughly 800KB XML file containing all text data (but not the separately-stored images or multimedia files) must be loaded entirely before the project is viewed, rendering it unwieldy for slow connections if *RolandHT* is accessed over the web. However, once it is loaded, subsequent operations are performed client-side and connection speed no longer matters, except when fetching the aforementioned images and multimedia files.

The notion of corpus

The notion of *corpus* is used in the study of languages (corpus linguistics), literature, and objects (architecture, archaeology). Definitions of _*corpus*_ provided by several different dictionaries of English and of literary terms are variations of "a related 'body' of writings, usually sharing the same author or subject-matter." (Baldick 52) It is almost never further defined in the context of specific studies. Instead, *corpus* is taken for granted, and is found far more often in headings than in discussion. When it is used in the text itself, the set of objects under consideration invariably has an author, time period

and/or geographical area in common. None of these things can usefully delimit the Roland corpus. However, some past uses of *corpus* would justify the use of the word to describe works about Roland.

Most often the contested word is used in corpus linguistics. One book in particular defines *corpus* as "a subset of an ETL [electronic text library], built according to explicit design criteria for a specific purpose." (Ghadessy et al. 179) This is the only non-dictionary definition I have found, and if we substitute "Western artistic production" for "ETL," it becomes relevant to the construction of *RolandHT*. The purpose of defining it is to shed light on how its protagonist has survived so unusually long while at the same time undergoing fundamental changes. However, the subjective nature of the definition process demands a willingness to alter design criteria as texts are encoded, and a leap of faith that scholars may be reluctant to make. Working with a semantically encoded body of work requires acceptance of XML as a valid format for making a scholarly argument. This in turn calls for critical responses that are also at least partly expressed in XML. ("I disagree with this encoding. How about this alternative instead?")

Roland is a flexible corpus that emerges and evolves in the course of its analysis. Its unifying subject may be a default connecting thread by virtue of Roland's name, but since the actual connecting threads are recurrent themes and imagery, the way to ascertain a work's status as a corpus object (or not) is a close reading of it. In this case, micro-results of the close reading are recorded by XML encoding.

The above describes a practical application of Gregory Ulmer's theory of heuretics. In his 1994 *Heuretics: The Logic of Invention*, Ulmer states that theory is assimilated into the humanities in two principal ways: by critical interpretation and by artistic experiment. Heuretics is the latter – it is a heuristic approach to theory, a reading process that, instead of attempting to theorize "what might be the meaning of an existing work," guides "a generative experiment: Based on a given theory, how might another text be composed?" (5)

Why not use words that have been used more frequently to describe Roland, such as *myth* or *legend*? Both of these terms imply homogeneity that the corpus does not possess, and deny Roland's historicity, however tenuous. Similarly, the terms closest to *corpus* – *oeuvre* and *canon* – are limiting to the point of inaccuracy: the Roland corpus has too many authors to be called oeuvre and too little institutionalized authority to be a canon.

Finally, *RolandHT* is not an archive. It does not strive for completeness, either within single works (only excerpts are presented) or within the corpus (which is, to date, representative but incomplete). More importantly, *RolandHT* is intended to be the opposite of a static collection of unchanging documents; its primary-source contents, encoding and annotations are all meant to change as research progresses.

Summary

If semantic encoding is to be taken seriously as a research tool by philologists, art historians and other humanists used to working in more traditional ways, a common language must first be built and justified. At Digital Humanities 2007 I hope to contribute to the effort of expanding the inflluence of digital research methods into mainstream humanities. I will do this by expanding the notion of *corpus* to include bodies of works that, because of their disparate formats, do not lend themselves to formal study using traditional methods.

---

1. The electronic part of the dissertation can be found at `<http://wordsend.org/rht/xml/index_2006.php>`. N.B.: it is a work in progress. New content is being added often. Please note also that the project can only be viewed using Firefox/Mozilla or another XSLT-aware browser.

Bibliography

Ariosto, Ludovico. Ed. Rudolf Gottfried. *Orlando Furioso; Selections from the Translation of Sir John Harrington.* Bloomington, IN: Indiana University Press, 1963.

Beowulf. Dir. Graham Baker, Perfs. Christopher Lambert, Rhona Mitra, Oliver Cotton and Götz Otto. DVD. Panorama Entertainment, 1999.

Baldick, Chris. *The Concise Oxford Dictionary of Literary Terms.* Oxford: Oxford University Press, 2004.

Boiardo, Matteo Maria. *Orlando Innamorato.* Trans. Charles Stanley Ross. Berkley, CA: University of California Press, 1989.

Browning, Robert. "Childe Roland to the Dark Tower Came." English Poetry Full-Text Database. Cambridge, UK: Chadwick-Healey Ltd. Accessed 2004-03-26.

Ghadessy, Mohsen, Alex Henry, and Robert L. Roseberry, eds. *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam: J. Benjamins Pub. Co., 2001.

King, Stephen. *Dark Tower I-VII.* New York: Signet, 1989-2003.

Konrad, Der Pfaffe. *Priest Konrad's Song of Roland.* Trans. J. W. Thomas. Columbia, SC: Camden House, 1994.

Roach, Greg. *Madness of Roland.* CD-ROM. Los Angeles, CA: HyperBole Studios, 1995.

The Song of Roland. Trans. Dorothy L. Sayers. Harmondsworth, UK: Penguin Books, 1957.

Silverstein, Shel. "Roland the Roadie And Gertrude the Groupie." *Belly Up!.* Perf. Dr. Hook and the Medicine Show. CBS, 1973.

Ulmer, Gregory L. *Heuretics: The Logic of Invention.* Baltimore, MD: Johns Hopkins University Press, 1994.

Применение технологий TEI и OAI в информационно-аналитической системе «Манускрипт»

Вотинцев Павел Анатольевич *(pvl@udsu.ru)*

Удмуртский государственный университет, Ижевск, Россия

В настоящее время по крайней мере несколько сотен древнейших славянских письменных памятников XI-XIV вв., имеющих большую научную, культурную, историческую ценность, недоступны широкому кругу исследователей разных специальностей и всем, кто интересуется русской историей и культурой. Причины просты: большая часть рукописей еще не издана, а если и есть издание, то оно является раритетом и работа с ним возможна только в крупных центральных библиотеках. Работа же непосредственно с рукописью часто осложняется удаленностью места работу исследователя от места хранения памятника. Но даже издание памятника в ряде случаев требует постоянного обращения к оригиналу, если дело касается его палеографии, оформления и других аналогичных вопросов. Кроме того, в существующих печатных изданиях, к сожалению, часто отсутствуют необходимые для знакомства с текстом и для его исследования справочные материалы, а сами рукописи изданы со значительным упрощением оригинала. Выходом из существующей ситуации недоступности письменных памятников древности и средневековья для научной, учебной и популяризаторской деятельности является создание электронных публикаций, которые способны не только заменить печатные издания, но и во многом перевести работу с ними на функционально иной уровень: (a) предоставить постоянный доступ к электронным копиям рукописей и к справочным материалом через интернет, (b) предоставить пользователю возможность выборки данных на основе многофункциональных справочных интерфейсов как на уровне мета- и аналитических данных, так и на уровне собственно грамматических, семантических, тематических характеристик текстов, (c) дать возможность коллективной работы над изучением памятника.

Одним из примеров информационной системы, предназначенной для решения части указанных

проблем, является информационно-аналитическая система «Манускрипт» (ИАС «Манускрипт»)

ИАС «Манускрипт» ориентирована на ввод, хранение, автоматизированную обработку, исследование и издание древних текстов, имеющих идентичные оригиналу графико-орфографический вид и структуру. В настоящее время ИАС «Манускрипт» уже содержит несколько десятков уникальных древних и средневековых славянских текстов.

В настоящее время ИАС «Манускрипт» состоит из нескольких разнофункциональных модулей, взаимодействующих с единой базой данных, в которой хранятся иерархически структурированные данные:

- специализированный редактор,
- модуль запросов и выборок,
- модуль электронных изданий,
- модуль лингвистических словарей,
- модуль выгрузки текстов и справочных материалов для печатных публикаций.

Основной задачей коллектива авторов и разработчиков на данном этапе развития системы является предоставление накопленной информации для анализа и изучения широкому кругу исследователей. Основной путь – создание специализированных модулей, предназначенных для и функционально разнообразного, дружественного и понятного манипулирования данными. Второй задачей является создание средств обмена мета-, аналитическими и текстовыми данными между коллективами, которые создают аналогичные по материалу коллекции, с целью предоставления пользователям возможностей использовать для анализа рукописей и текстов инструментальные средства работы с древними документами, созданные различными коллективами.

Для обеспечения максимально широкого доступа различными инструментальными средствами к накопленной в ИАС информации необходимо использовать стандартизованные форматы и средства обмена данными.

В качестве основы для решения этой задачи решено использовать технологии Open Archive Initiative (OAI) и Text Encoding Initiative (TEI):

- OAI – для индексации существующих текстов и их фрагментов;
- TEI – для предоставления мета и аналитической информации, а также самих текстов.

В настоящее время ведутся работы по включению ИАС «Манускрипт» в состав открытого сообщества OAI в качестве Data Provider. Для предоставления метаинформации в формате Dublin Core разрабатывается отдельный модуль системы. Данный модуль позволит собирать метаинформацию не только по самим текстам, но и по их фрагментам (различные элементы иерархий).

Найденный уникальный идентификатор средствами интерфейса OAI возможно использовать для получения метаинформации или полного текста из системы «Манускрипт» в формате XML-TEI.

Сегодня рекомендации консорциума TEI являются общепризнанными в мире. Поэтому вполне логичным стало решение разработчиков ИАС «Манускрипт» использовать именно этот инструмент для описания метаинформации и полных документов для обмена данными.

ИАС «Манускрипт» использует сетевую модель данных для описания объектов (единиц) рукописей и связей между ними, поэтому при создании записей в формате XML-TEI основной проблемой стали структурные ограничения, накладываемые стандартами XML: структура XML документа, которым является документ в формате XML-TEI, должна быть строго иерархической. Следствием этого, например, является невозможность получить всю информацию о тексте (имеющуюся в системе «Манускрипт») в виде одного документа.

Для реализации интегрирования технологий ИАС «Манускрипт» с технологиями TEI разрабатываются два модуля.

Модуль хранения и обработки мета- и аналитических данных предназначен для работы с археографическими, текстологическими и иными сведениями о самих рукописях, текстах и их фрагментах, а также для формирования запросов и выборок на основе этих сведений. Средствами этого модуля поиск возможен только по мета- и аналитическим данным.

Запросы пользователей на получение определенных элементов текста в формате XML-TEI из системы «Манускрипт» обрабатывает специальный модуль. Основные возможности данного модуля:

- получение любого элемента текста в формате XML-TEI, описанного в ИАС «Манускрипт» в качестве самостоятельной единицы;
- расстановка приоритетов по созданию документов XML-TEI. Данная опция необходима ввиду накладываемых ограничений на структуру XML документа. Пользователь сам может выбирать, какая структура для него является приоритетной, а какая – менее приоритетной. Иначе говоря, при создании документа максимально отображается приоритетная структура, внутри нее отображается следующая по приоритету и т.д.;

- получение мета и аналитической информации как вместе с полным текстом, так и отдельно;
- запрашиваемый документ собирается в момент запроса, поэтому любое изменение в тексте тут же будет использовано.

ИАС «Манускрипт» постоянно развивается, и вместе с ней модернизируются указанные инструменты.

Во время доклада будет продемонстрирована работа ИАС «Манускрипт», описанных модулей и технологий.

Index of Presenters

Index of Topic Keywords