Research Practice and Research Libraries: Working toward High-Impact Information Services

Carole L. Palmer
Center for Informatics Research in Science & Scholarship (CIRSS)

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

OCLC Programs and Research
19 June 2008

GSLiS

The problem in a nutshell…

Utopian e-research scenarios promoted decades ago may now be obtainable goals.

They will be enabled by the interplay of *technology* and *user behavior*.

- We have a reasonable understanding of changing technology but a limited understanding of changing user behavior … and therefore a poor understanding of the interplay
    - in the actual activities of reading, experimenting, analyzing, interpreting and problem solving.
- One problem is that much of our research doesn't identify the features most likely to be explanatory and predictive, or indicate what interventions can make a real difference.
- In what follows, I draw on our studies of scholarly information work over the past decade to discuss how information use is changing in the practice of science and scholarship and reflect on where research libraries can direct their efforts to make a significant contribution.

Higher stakes in getting information services right

In the contemporary context of e-science, aiming directly to re-shape scientific endeavours and provide new infrastructures to support them, [the] goal of studying the detail of actual practice takes on a new significance. (Hine, 2005)

The body of research on general trends in digital information use provides and important base, but often only a silhouette of the interplay between researchers and information.

Studies need to be refined to investigate the role and value of information and how to improve research.

- how information fits in, interacts, fuels new discoveries
- what differences make a difference: disciplines and domains, methodological strategies, project stages, etc.

The story line

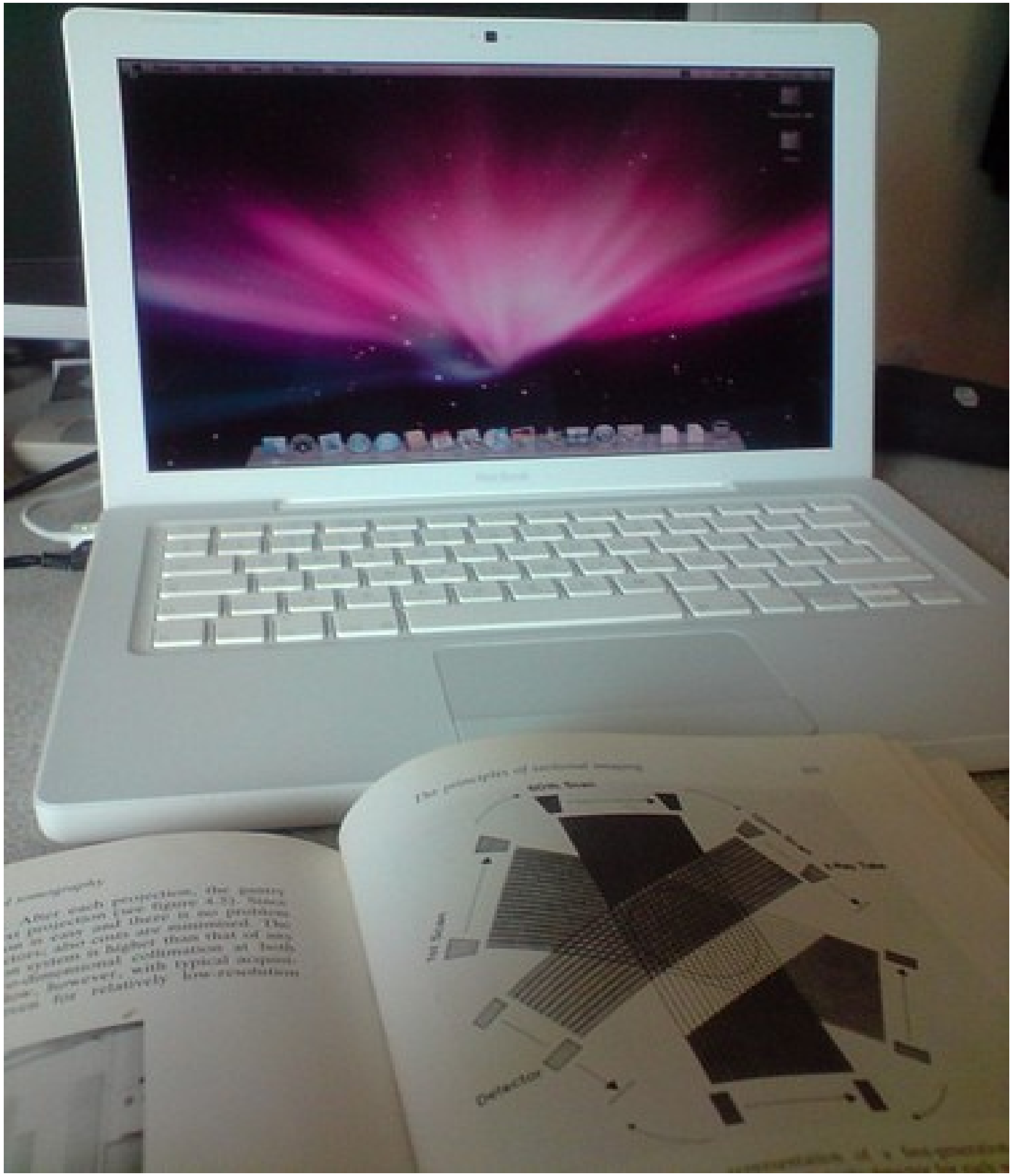
- We need to know more about scholarly research practices—how scholars are working & wish to work with information,
  - the case of “reading”

- and determine what kinds of information support can really make a difference in how scholars work.
  - insights from a study of scientific discovery

- Management and reuse of data sets is one such area that depends on deep understanding of research practice,
  - insights from research on federating cultural heritage collections

- and on readying research librarianship for data curation responsibilities
  - the need to step up, but with skepticism.

Reading is complex



General trends in e-journal use well documented

Nearly all STM journals are now available electronically

- access in the sciences is predominantly to these electronic versions
- 98% of medical researchers prefer e-journals (Hemminger, 2007)

Web “bouncing” common, especially in medicine, life sciences

(CIBER group - Nicholas, et al., 2006)

Number of articles read is rising

- over 30% higher in 2006 than in the mid-90s

Reading time per article is falling

- medical researchers about 24 minutes per article (Tenopir, 2006)

But are these really indicators of reading?

- Our studies suggest researchers are not reading more, but rather scanning, exploring, and getting exposure to more sources. (Palmer, 2001, 2002)

- Consistent with the recent reports by Tenopir and CIBER

- In fact, researchers may be practicing active reading avoidance. (Palmer, 2007; Renear, 2006, 2007)

  Researchers are rapidly navigating through more material, spending less and less time with each item, and attempting to assess and exploit content with as little actual reading as possible.

Intensification of longstanding practices

- Indexing and citations help us decide whether or not articles are relevant … without reading them.
- Abstracts and literature reviews help us take advantage of articles … without reading them.
- The articles we do read provide summaries and discussions that help us take advantage of other articles… without reading them.
- Colleagues, and graduate students, help us learn about and understand articles… without reading them.
- And the apparatus (tables of contents, references, figures, etc.), distinctive formatting of text components (such as lists, equations, scientific names, etc.), help us exploit articles … without reading them.

But researchers do “read”, in many different ways

- probing in new areas → conference lurking to web exploration
- learning → textbook-like explanations
- positioning → directed searching of topic
- competing → directed searching of people
- scanning, stay aware → reviews to alerting services & blogs
- rereading → personal collections
- reading around → following leads to thematic collections

Other uses of the literature are equally important

consulting - experimental resource to identify

- protocols
- instrumentation
- comparative results

compiling – customized personal collections

- laptops full of PDFs

extracting – core knowledge base

- “facts” for ontology development

building - source for database enrichment

- annotation, evidence

Supporting creative and indirect uses of the literature

Finding articles to read — left-to-right, top-to-bottom — is even less of an accurate representation of literature use than it ever was.

- We “read” less and less every year, yet are even more analytically engaged with the literature

But the value of functions are far from uniform across fields:

- In the humanities, reading around, collecting, and rereading
- In the sciences, researchers likely to benefit from fast-paced, indirect, “horizontal” use of the literature.

Advances dependent on

- encoding and associated metadata and ontologies
- greater application of analytical text mining and literature-based discovery

Scientific discovery is work

Flickr user: stancia creative commons

How do we improve conditions for discovery?

Information and Discovery in Neuroscience (IDN Project)
NSF/CISE/Digital Technologies and Society, #0222848

- What information conditions are associated with advancements and problems during the course of research?

- What role can literature based discovery (LBD) play in daily scientific practice?

- Partnership with Arrowsmith Project

  Based on Swanson's (1986) notion of "undiscovered public knowledge"

  Smalheiser & Swanson's system adapted for PubMed end users

  Conceived of as tool for hypothesis testing – implicit relationships among literature A and literature C.

Study of information practices <u>and</u> informatics efforts

12 project-based cases at 4 labs, 11 key informants, 25 total participants
1/3 of participants field testers for Arrowsmith

Qualitative Interviewing	(44 sessions)
• project-based	
• critical incidents	(progress, problems, shifts)
Information Diary	(137 records)
• Arrowsmith search logs	
• Information activity logs	
Field Observation	(19 hours)
• information activities	
• research processes	
• work environment	

Key aspects of research design

- Partnering with neuroscientists

  who are actively investing in and customizing digital resources and tools for themselves and their communities

  best indicators of how researchers wish to engage with information technology in their work.

- Longitudinal case study

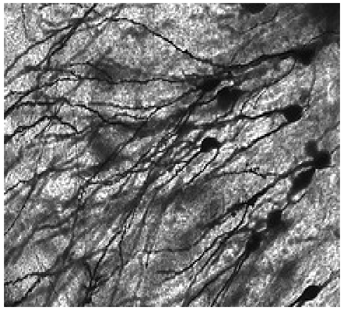
  chronicling of projects and relationship to larger programs of research

  extended use of personal diaries in conjunction with critical incident interview data

  verification of reported information activities and importance over time

  refinement and validation of our information categorization scheme

Rich cases representing range of neurosciences



	LAB 1	LAB 2	LAB 3
Research types / techniques	clinical studies and computational neuroscience - fMRI	neuronal substrate of learning and memory -electrophysiology	microscopy, telescience, and anatomy - microscopy and tomography
Project Characterizations	neuroinformatics - computing tools for neuroscience application clinical neuroscience - investigating reward systems using brain area activation	basic neuroscience – affect of lesions on acquisition and extinction of discriminative behavior	basic neuroscience - characterizing mouse models of disease (using microscopy and imaging techniques) ontology development for shared databases

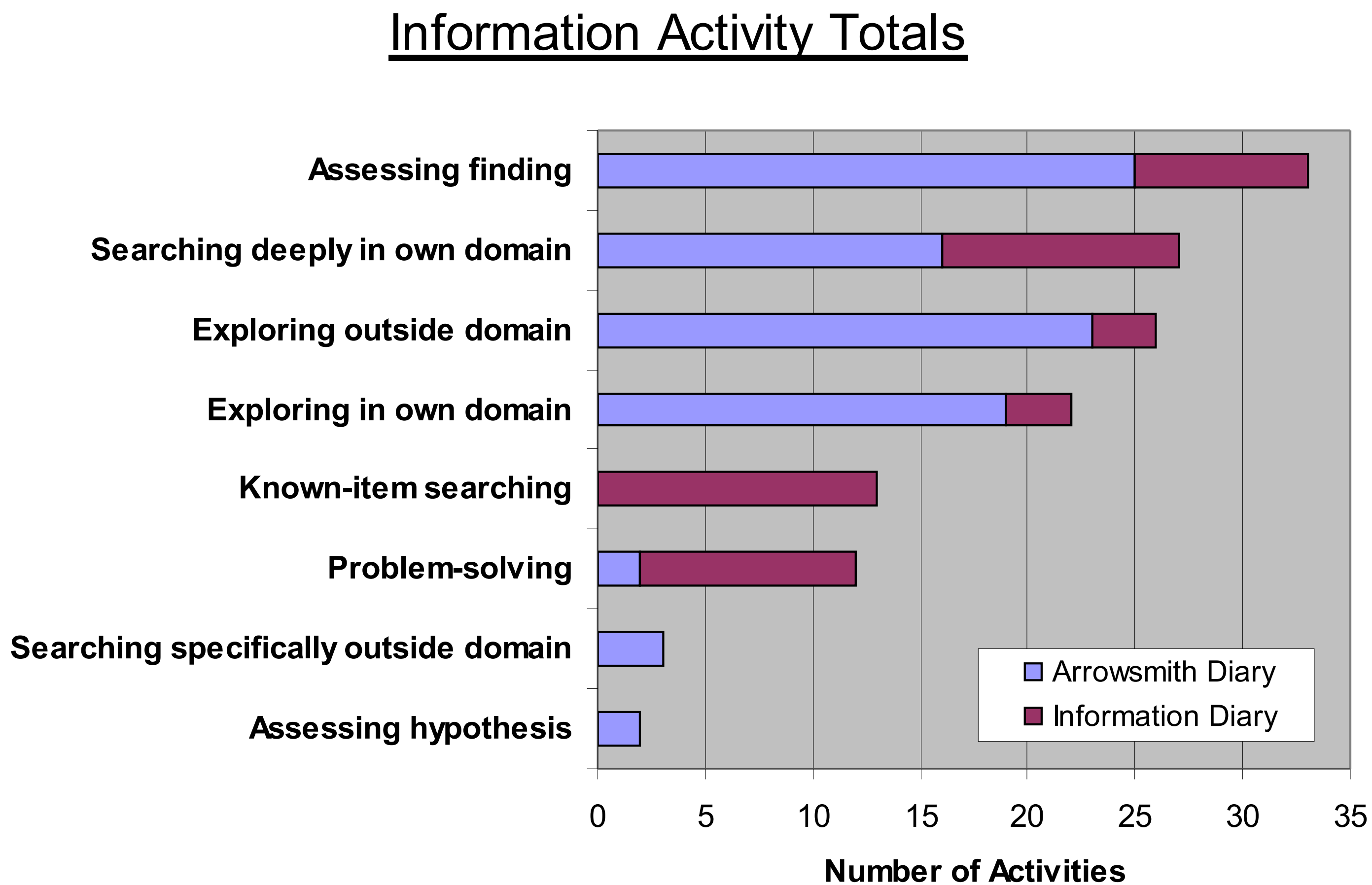
Primary Domains
 	
(as represented in collaborations and use of literature)
	- computer science
	- computational neuroscience
	    - modeling
	- imaging
	    - fMRI (functional, structural)
	- psychology
	- psychiatry
		- electrophysiology
		- behavioral neuroscience
		- anatomy
		- cell biology
		- biochemistry
		- neuropsychology
		- neurophysiology
			- anatomy
			- microscopy
			- computer science
			- biology
			- neuroinformatics
			- biochemistry
			- neurophysiology

Progress and problems related to information work

- Greatest advancements associated with visualization of data
- Knowledge of brain anatomy (people, information resources and tools) playing pivotal role in moving research forward
- Difficulty locating specifics on protocols, instrumentation, measurements, experimental context, etc.
- Retrospective, non-digital literature often ignored
- Review articles essential for keeping up with information and for learning in new areas

Unexpected LBD applications

Surprisingly, hypothesis assessment rare with Arrowsmith



Most frequent activities

Assessing finding against the literature

- How important is this result?
  increased in frequency over time

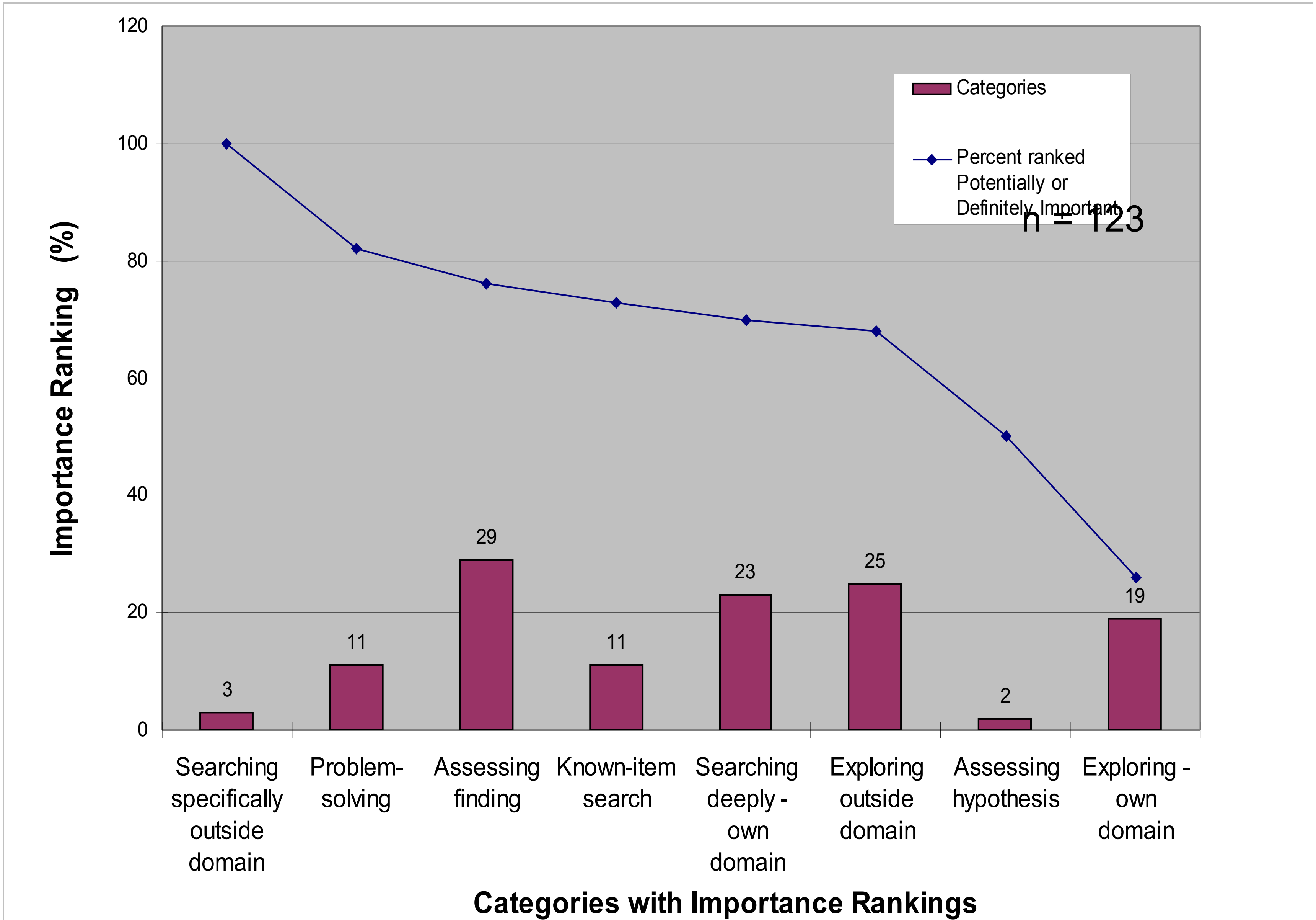
Exploring outside own domain

- What am I missing?
  54% focused on clinical concepts or diseases
  difficulty evaluating importance of information found

Searching deeply in own domain

- Is this project worth investing in?
  analyzing risk or verifying viability of a research project

But, low frequency more “important” for discovery



Information work as weak or strong

Extending Herbert Simon's conceptualization of weak / strong methods
(Simon, Langley, and Bradshaw, 1981)

Weak (novice, trial & error)	**Strong** (expert, tried-and-true)
Ill-structured problem space	Structured problem space
Unsystematic steps	Systematic steps
Low domain knowledge	High domain knowledge
Data driven	Theory driven
Seek and search	Recognize and calculate

Importance of weak approaches

- "... fundamentality of a piece of scientific work is almost inversely proportional to the clarity of vision with which it can be planned."
  (Simon, Langley, & Bradshaw, 1981, p. 5).

- may be all that is available on the frontiers of knowledge
  (Simon et al., 1987)

- required for revolutionary science (Kuhn, 1962)

And, our previous studies of interdisciplinary scientists and scholars show weak conditions common in their research.
(Palmer 1996, 1999, 2001; Palmer & Neumann, 2002)

How does the weak/strong framework help us?

Strong information work is most routine and codified
Weak information work is the most arduous and most speculative

Weak work highest in preparation stages of research
- Assessing preliminary hypotheses
- Feasibility assessment
- Building new interdisciplinary collaborations

High in all cases where new learning involved
- Developing a new research technique

The most productive points for information support are likely to be at ends of the weak / strong continuum.

Can predict the kinds of activities and stages of research where weak and strong information work will be centralized.

(Palmer, Cragin, & Hogan, 2007)

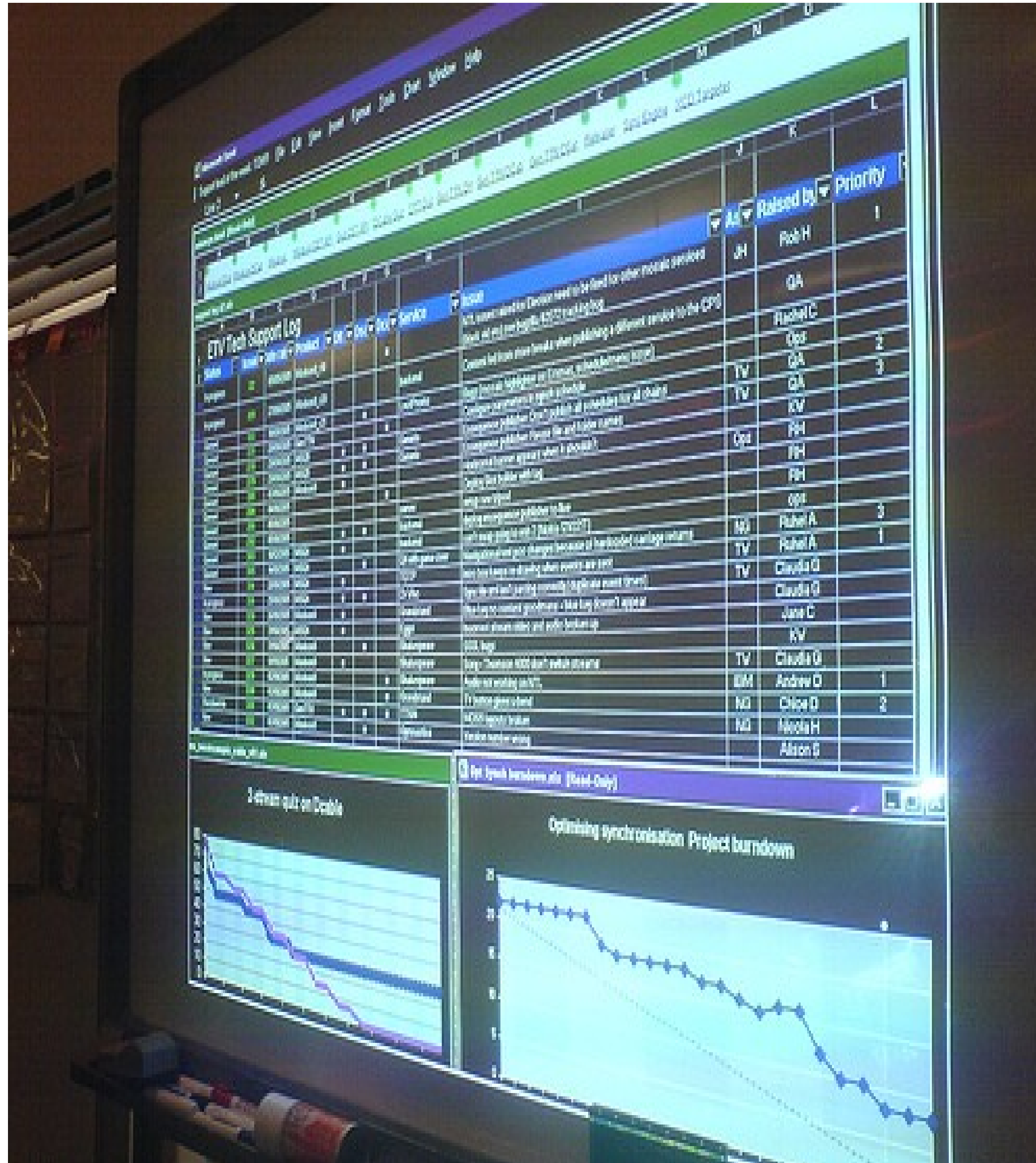
Strengthening weak work

Some, but not all, weak work should be stronger, more routine, codified,

especially in informatics and data intensive research

- literature based discovery for hypothesis testing
- instrumentation and methods fact-finding
- ontology and standards development for data repositories
- management and reuse of data

Data sets
as
“special
Collections”



Flickr `: r h creative commons

Curation Profiles Project (IMLS NLG 2007-2009)

CIRSS with Purdue University Libraries (D. Scott Brandt, PI)

- Investigating curation requirements across sciences
- in collaboration with librarians working closely with researchers on issues of scientific research data management and curation
    - researcher data / metadata workflow
    - policies for archiving and access
    - system requirements for managing data in a repository
    - identify roles of librarians and skill sets they need to support archiving and sharing

Complexities of data collections

- Primary and secondary data, mobilized to produce new primary research, and their various transformations
- Generated by instruments, people, in the lab, in the field, etc.

  - data characteristics
  - storage & security
  - standards / metadata / interoperability
  - preservation
  - access
  - sharing
  - intellectual property
  - quality control
  - services
    - linking & citation
    - visualization

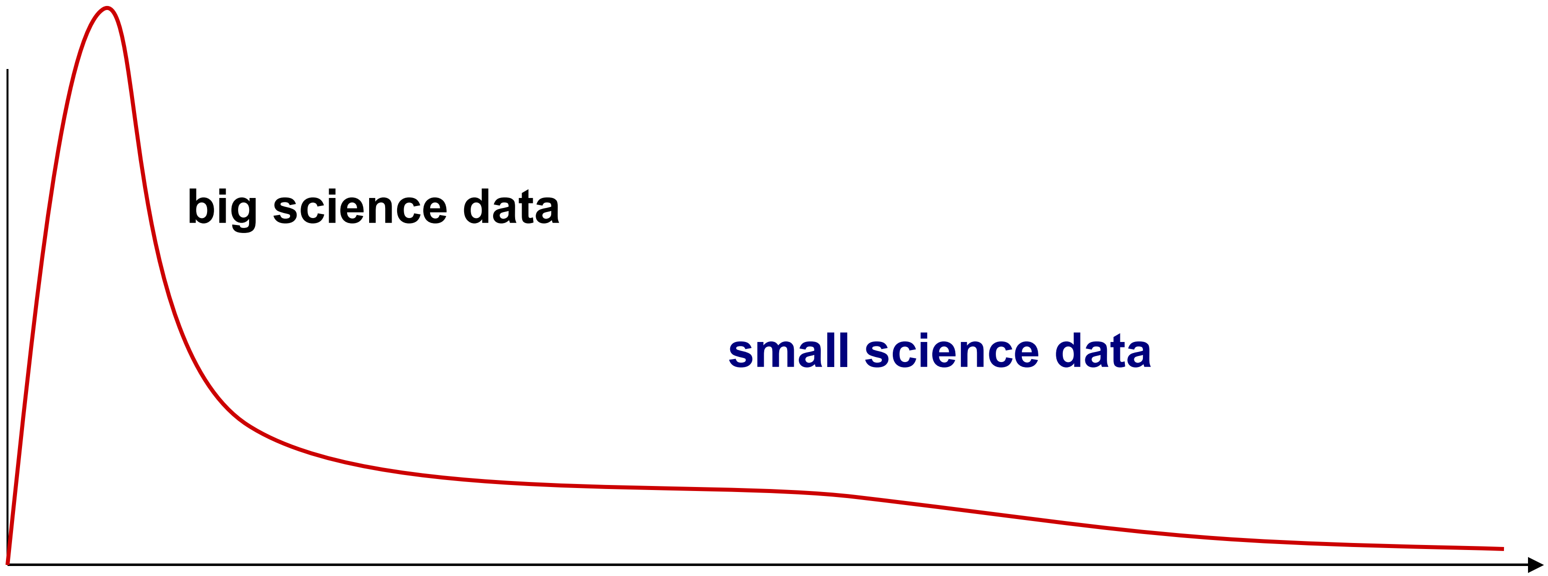
Data Characteristics	**Crystallography**
Type	1. "Raw data" – binary image frames 2. "Phased file" – electron density 3. "Integrated data" – amplitudes of molecules 4. "Corrected data" – according to theory
Format	1. Binary diffraction images based on the software 2. Different electron density image 3. Multiple formats 4. CIF file
Size	1. About 2,400 frames ¼ -1Mb each – about/over 1Gb 2. > 100Mb 3. 5-6 Mb 4. < 1 Mb
Workflow	well-defined stages, for measurement or analytical purposes, in sequence; output of one stage constitutes the input to the next; for publication CIF considered final result of experiment

Research libraries' role most evident in small science

Data from Big Science is … easier to handle, understand and archive.

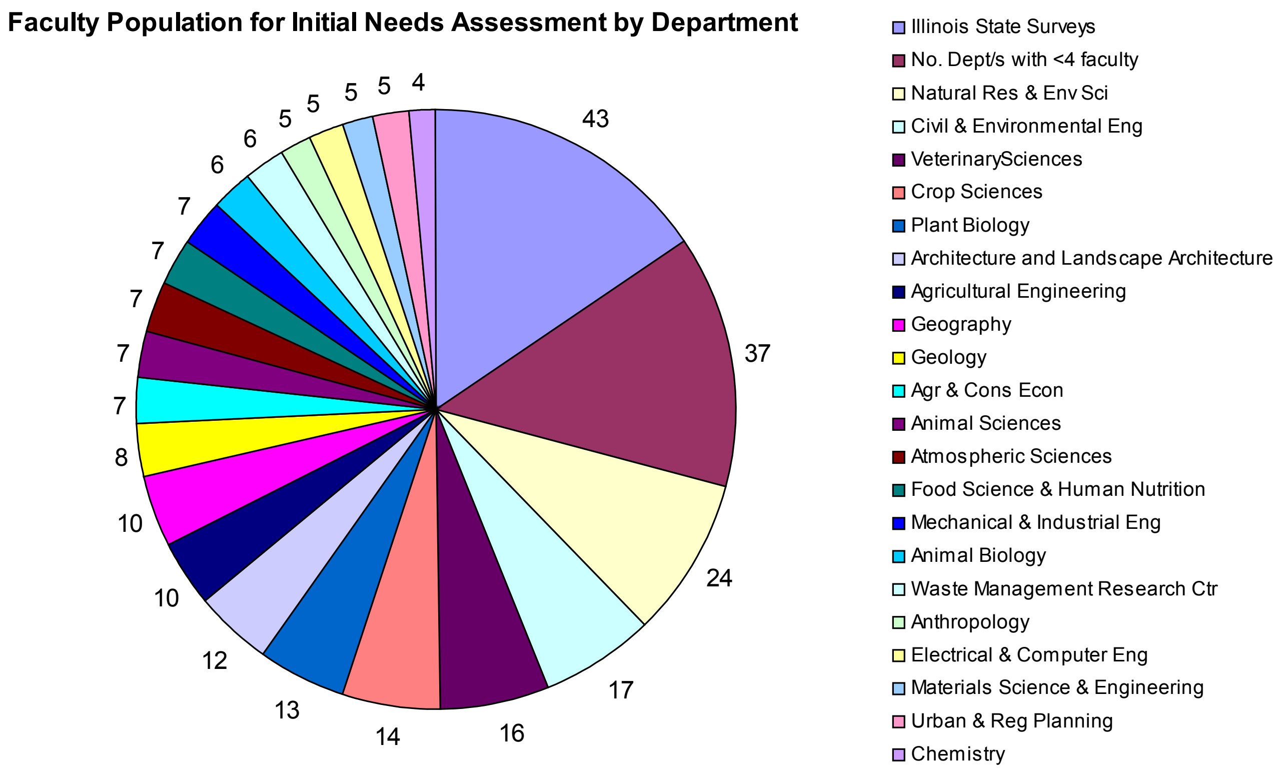
Small Science is horribly heterogeneous and far more vast. In time Small Science will generate 2-3 times more data than Big Science.

('Lost in a Sea of Science Data' S.Carlson, The Chronicle of Higher Education, 23/06/2006.)



Challenges of small, cross-disciplinary science

Data needs assessment of UIUC “Faculty of the Environment”; daunting to define, reach, respond to the user community.



How do we identify and represent "analytical potential"

Researchers have clear ideas about what data sets do *not* need to be saved or preserved, but may not be able to predict potential of

- long-term use by others, especially for applications in other fields
- collective value or applications of the many, often specialized, distributed collections in large-scale aggregations
  - theoretical modelers earliest adopters

With cultural heritage collections, decades of opportunity-driven digital "projects" have resulted in overall lack of cohesion of digital content.

- Need to aim for contextual mass, not just critical mass (Palmer, 2004) through more systematic collection of complementary content

What are the meaningful organizing units for data sets?

Fundamental problems of scale & granularity

- Flat representation of digital collections; small window into large, diverse accumulation of content
  - all items appear equal
  - strengths, special features not evident

- Diminished “intentionality”
  - purpose of and relationships among collections not evident

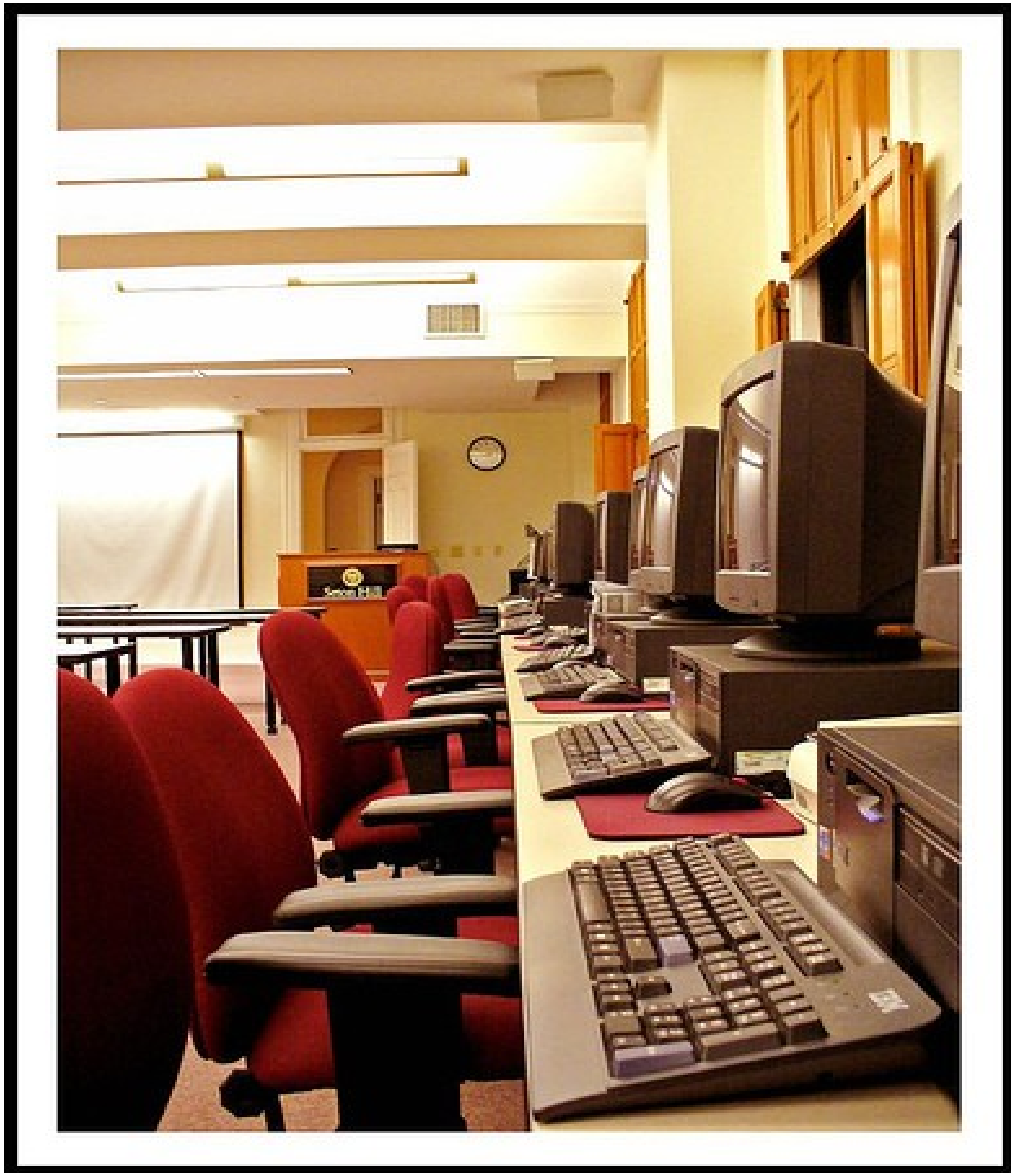
- Collection level metadata solutions not straightforward
  - what constitutes a set
  - how to handle transformations and new composites, and relationships to original sets

Data

curation

is

contentious



K. Sawyer creative commons

What does LIS have to offer data curation?

In the tradition of research librarianship, professionals must understand the landscape of research resources and how resources work together:

- Collect and manage data in ways that add value

  and

  promote sharing and integration across laboratories, institutions, and fields of research.

- Build and maintain data systems that work in concert with digital libraries, archives, and repositories,

  and

  the indexing systems, metadata standards, ontologies, etc. associated with digital data and products.

Extending library functions to new content

The active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education.

Activities

- enable data discovery and retrieval
- maintain data quality
- add value
- provide for re-use over time
- archiving
- preservation

Tasks

- appraisal and selection
- representation
- authentication
- data integrity
- maintaining links
- format conversions

What's new for libraries and librarians?

- Closer engagement with scientists during research production,
  - more sophisticated understanding of the differences in research cultures across domains
  - potential for more direct contributions to the scientific enterprise

- Facilitation of data deposition to
  - local, disciplinary, larger federations

- New collaborations and constituencies
  - campus IT, research officers

- Development of data curation principles and systematic practices

Professionalizing curation of research data

CIRSS initiatives with research / data centers in the sciences and humanities to develop

- Data curation concentration in MSLIS

2 IMLS – Laura Bush 21st Century Librarian Program Grants

Science, Heidorn, PI / Humanities, Renear, PI

Focus on digital data collection and management, representation, preservation, archiving, standards, and policy.

Develop curriculum, internships, promote & share DC expertise.

- 1st summer institute for academic librarians, June 2008
- Digital Curation Centre's 6th International Conference in 2010

Curators inside research libraries & research centers

Science Partners

- Biomedical Informatics Research Network (BIRN), UCSD
- Missouri Botanical Garden
- Smithsonian Institution
- Field Museum of Natural History
- U.S. Geological Survey
- Marine Biological Laboratory
- US Army ERDC-CERL

Humanities Partners

- Institute for Technology in the Arts and Humanities (IATH),
- Committee on Documentation (CIDOC) of the International Council of Museums (ICOM)
- Center for Computing in the Humanities, Kings College London
- OCLC
- Women Writers Project
- Perseus

References

Hemminger, B. M., Lu, D., Vaughan, K.T.L., Adams, S. J. (in press). Information seeking behavior of academic scientists. Journal of the American Society for Information Science & Technology.

Hine, C. (2005). Material culture and the shaping of e-science. First International Conference on E-Social Science. Manchester, UK. http://www.ncess.ac.uk/events/conference/2005/papers/papers/ncess2005_paper_Hine.pdf.

Nicholas, D., Huntington, P., Jamali, H. R., & Dobrowolski, T. (2006). Characterising and evaluating information seeking behaviour in a digital environment: Spotlight on the 'bouncer'. Information Processing and Management 43, 1085-1102.

Palmer, C. L. (1996). Information work at the boundaries of science: Linking information services to research practices. Library Trends 45(2), 165-191.

Palmer, C. L. (1999). Structures and strategies of interdisciplinary science. Journal of the American Society for Information Science 50(3), 242-253.

Palmer, C. L. (2001). Work at the Boundaries of Science: Information and the Interdisciplinary Research Process. Dordrecht: Kluwer.

Palmer, C. L. & Neumann, L. (2002). The information work of interdisciplinary humanities scholars: Exploration and translation. Library Quarterly 72 (January), 85-117.

Palmer, C. L., Cragin, M. H., and Hogan, T.P. (2007). Weak information work in scientific discovery. Information Processing and Mangement 43 no. 3: 808-820.

Renear, A. H. (2006). Ontologies and STM publishing. STM Innovations, London, UK, 1 December, 2006.

Renear, A. H. (2007). Standard domain ontologies: The rate limiting step for the "Next Big Change" in scientific communication.
The 233rd American Chemical Society National Meeting, Chicago, IL, 25-29 March, 2007.

Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. Synthese, 47(1), 1-27.

Swanson, D.R. (1986). Undiscovered public knowledge. Library Quarterly, 56(2), 103-18.

Tenopir, C. (2006). How electronic journals are changing scholarly reading patterns. CONCERT Annual Meeting, Taipei, Taiwan, 2006.

Questions & comments, please

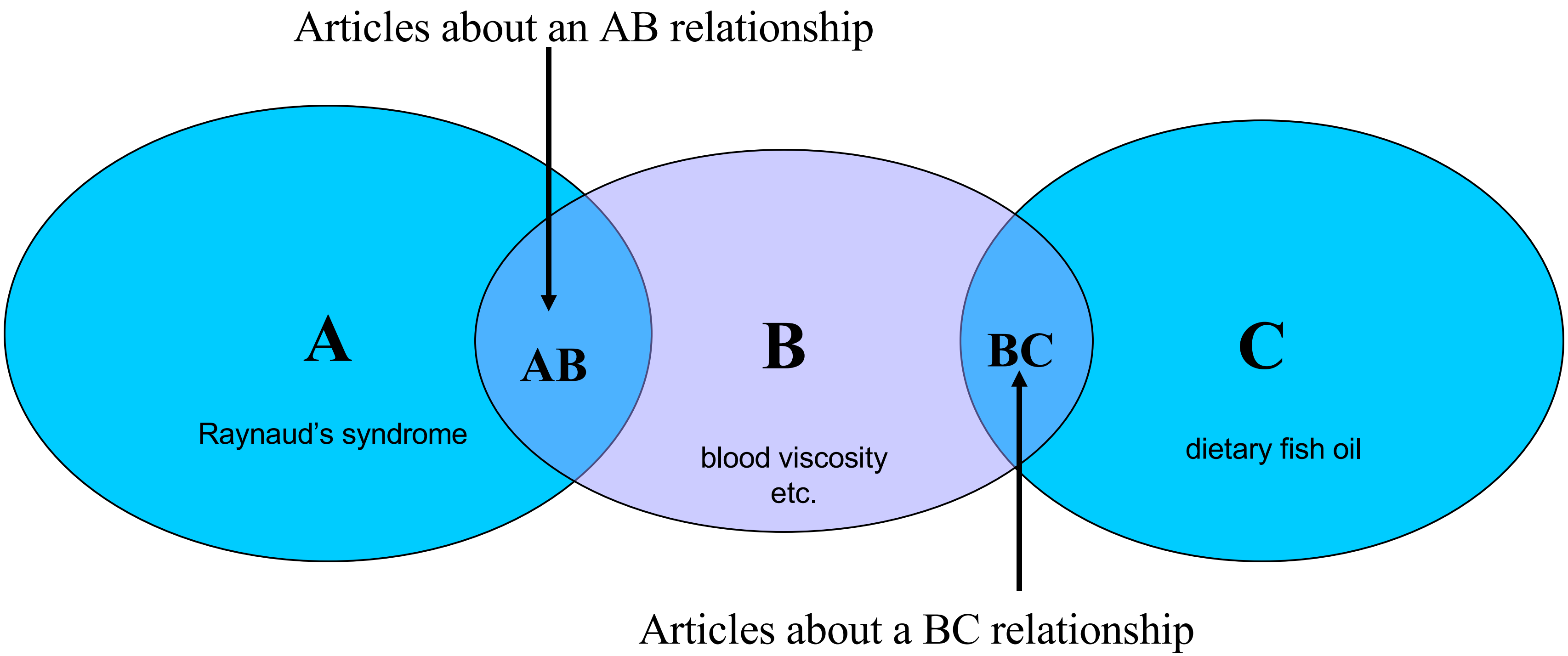
clpalmer@illinois.edu

Center for Informatics Research in Science and Scholarship (CIRSS)

http://cirss.lis.uiuc.edu/

CIRSS

Arrowsmith LBD: the ABC Model



- AB and BC are complementary but disjoint : They can reveal an implicit relationship between A and C in the absence of any explicit relation.
- The researcher assesses titles in the B literature identified by the system for fit or contribution to problem.