Adapting Digital Information to Scientific Practices

Carole L. Palmer

Center for Informatics Research in Science and Scholarship (CIRSS)

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

STM Spring Conference
The Next Generation: Endless Choices & Economic Constraints
Cambridge, MA, 24-26 April 2007
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.
  - Both of which are changing rapidly.
- Yet while we have a reasonable understanding of changing technology, we have a poor understanding of changing user behavior ... and therefore a poor understanding of the interplay.
- One problem is that the most commonly applied research methods don’t identify the features most likely to be explanatory and predictive, and which would present opportunities for intervention.
  - For instance, measures of accessing, downloading, linking, and citing behavior tell us little about
    - the actual activities of reading, experimenting, and problem solving,
    - or about the features of information that fuel new discoveries,
    - or how to predict changing behavior and response to new technologies and practices.
- In what follows, I discuss different dimensions of information use in the practice of science and how they are changing.
  - This use-centric view is based on results from our studies of information use in varying modes of interdisciplinary science, our model of “weak” and “strong” information work in the research process, and field tests of a literature-based discovery (LBD) tool in neuroscience laboratories.
  - This gives a sense of the sort of understanding more sensitive methods can provide.
- But there is a lot more to do.
Changing patterns in scientific information use

**Journals**

All STM journals are now available electronically

- Access is predominantly to these electronic versions
  - 98% of medical researchers prefer e-journals (Hemminger, forthcoming)

Number of articles read is rising

- Reading time per article is falling
  - For medical researchers about 24 minutes per article (Tenopir, 2006)

Web “bouncing” common, especially in medicine, life sciences (Ciber - Nicholas, et al., 2006)

**Other sources**

Increasing use of non-journal content, i.e. GenBank, lab research web pages

- Less dependence on personal communication (Hemminger, forthcoming)

Researchers prefer publisher over institutional repositories for finding research information (Pryor, StORe Project, 2006)
The body of research on general adoption of digital resources and trends in use provides only a silhouette of the interplay between scientists and information. We have accumulated many important findings from surveys and log analyses, but they don’t tell us enough about the role and value of information in research.

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)

In our research, we want to know how digital technologies can improve science.
• how information fits in, interacts, fuels what researchers do,
• what has the most potential for advancing research programs,
• what differences make a difference

-- disciplines and specializations, methods, data, stages of research, modes of collaboration, etc.

•
A closer look at “reading”

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

In fact, researchers may be practicing active reading avoidance  
(Rnear, 2006, 2007)

This is consistent with the recent reports by Tenopir and Ciber.

But, do these practices improve scientists’ ability to solve research problems, and if so, how?

So, the question of whether scientists are reading more or less is interesting, but other questions get us further.

What’s more important? Helping researchers find things to read or supporting better scanning and exploring?
What have we learned about reading practices?

Scientists read to:

- **probe** in new domains → web exploration
- **learn**, get oriented → location of textbook-like explanations
- **position** → directed searching of topic
- **compete** → directed searching of people
- **scan** the environment, stay aware → review of sources

Accesses, downloads, and citation don’t accurately represent any of these dimensions of reading. Indicators of interest, perhaps, but not actual use or value to science.
Scientists are not only readers

Other activities with scientific literature throughout the research process:

• **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
Scientists are not just seekers and consumers

**Authors**
- Standing, audience, speed of publication - top 3 factors in choosing journal publisher  
  (Hemminger, forthcoming)

- Faculty note that “visibility” and “tiered prestige” will never be duplicated in institutional repositories  
  (Palmer, IMIRD Project, 2007)

**Digital resource / tool developers**
Scientists are actively developing digital resources for themselves and their communities: data repositories, collaboratories, mining and analysis tools

Working to mobilize the specialized part of the digital realm they care about and know best. They are investing, customizing for practice.

These activities are the **best indicators** of how researchers wish to engage with information technology in their work.

Examining and collaborating with informatics tool builders tells us more about how to improve scientific practice than just looking at use of existing information resources or research practice in a general.
Scientific Communication Initiative (SCI)

Selected partners in research, development, and education
- Biomedical Informatics Research Network - UCSD
- Arrowsmith literature mining project, University of Illinois at Chicago, Neuroscience
- Smithsonian Institution
- American Museum of Natural History
- Missouri Botanical Garden

Selected projects include:
- Information and Discovery in Neuroscience, Palmer (NSF 0222848)
- HerbIS: Recorded Botanical Information Synthesizer, Heidorn (NSF DBI-0345341)
- Information specialist programs:
  - Biological informatics and Data Curation (NSF 0534567 and IMLS 05-06-0036-06)
- Researcher Browsing Behavior in Digital Environments – Palmer & Renear
- Informatics to Support Interdisciplinary Analysis of Alternative Land Use - Heidorn
- Networked Environmental Sonic Toolkits for Experimental Research - Downie
- Formal Semantics for XML Representations of Scientific Ontologies – Renear
Case studies of information practices and informatics efforts in brain research

Personnel: Carole Palmer
Melissa Cragin and Tim Hogan, Doctoral research assistants

Initiative: Scientific Communication Initiative, GSLIS, UIUC
http://sci.lis.uiuc.edu/

Support: NSF, Computer and Information Science and Engineering / Digital Technologies and Society - Grant No. 0222848

Previous work: Builds on study of interdisciplinary biological, physical, and behavioral scientists (Palmer 1996, 1999, 2001)
Research questions

In the case of contemporary brain research

• 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?

Partnered with Arrowsmith Project – Neil Smalheiser & Don Swanson’s system adapted for PubMed.

Based on Swanson’s (1986) notion of “undiscovered public knowledge”.
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.
Field testing of LBD in neuroscience labs

Studied 4 labs, 12 project cases, 11 key informants, 25 scientists total

Qualitative Interviewing (44 sessions, 60-90 minutes)
- project-based
- critical incidents - progress, problems, shifts

Information Diaries (137 records)
(interpretations validated through interviews)
- Arrowsmith search diaries
- Information activity diaries

Field Observation (19 hours)
- information activities
- research processes
- work environment

Document Analysis
Rich cases representing range of neurosciences

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

<table>
<thead>
<tr>
<th>Project Characterizations</th>
<th>LAB 1</th>
<th>LAB 2</th>
<th>LAB 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>- computer science</td>
<td>- electrophysiology</td>
<td>- anatomy</td>
<td></td>
</tr>
<tr>
<td>- computational neuroscience</td>
<td>- behavioral neuroscience</td>
<td>- microscopy</td>
<td></td>
</tr>
<tr>
<td>- modeling</td>
<td>- anatomy</td>
<td>- computer science</td>
<td></td>
</tr>
<tr>
<td>- imaging</td>
<td>- cell biology</td>
<td>- biology</td>
<td></td>
</tr>
<tr>
<td>- fMRI (functional, structural)</td>
<td>- biochemistry</td>
<td>- neuroinformatics</td>
<td></td>
</tr>
<tr>
<td>- psychology</td>
<td>- neuropsychology</td>
<td>- biochemistry</td>
<td></td>
</tr>
<tr>
<td>- psychiatry</td>
<td>- neurophysiology</td>
<td>- neurophysiology</td>
<td></td>
</tr>
</tbody>
</table>
Longitudinal case study advantages

- Chronicling of research projects and relationship to larger programs of research

- Extended use of personal diaries in conjunction with interview data
  
  critical incidents for interviews

  verification of reported information activities and their importance over time

  refinement and validation of our information activity categorization scheme
Unexpected LBD applications

Surprisingly, hypothesis assessment rare with Arrowsmith.
High frequency activities

Assessment of finding against the literature
  • increased in frequency over time

Exploring outside one’s own domain
  • 54% focused on clinical concepts or diseases
  • difficulty evaluating importance of information found

Searching deeply in one’s own domain
  • wide variety of aims, including looking for “links,” writing, and learning
  • analyzing risk or verifying viability of a research project
Most important activities

<table>
<thead>
<tr>
<th>Categories with Importance Rankings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance Ranking (%):</td>
</tr>
<tr>
<td>n = 123</td>
</tr>
</tbody>
</table>

- Searching specifically outside domain: 3
- Problemsolving: 11
- Assessing finding: 29
- Known-item search: 11
- Searching deeply - own domain: 23
- Exploring outside domain: 25
- Assessing hypothesis: 2
- Exploring - own domain: 19

Importance of Information Resulting from Activities

Searching specifically outside domain
- Problem-solving
- Assessing finding
- Known-item search
- Searching deeply - own domain
- Exploring outside domain
- Assessing hypothesis
- Exploring - own domain

Assessing hypothesis
- Exploring - own domain

Problem-solving
- Assessing hypothesis
- Exploring - own domain

Assessing finding
- Known-item search
- Searching deeply - own domain
- Exploring outside domain
- Assessing hypothesis
- Exploring - own domain

Known-item search
- Searching deeply - own domain
- Exploring outside domain
- Assessing hypothesis
- Exploring - own domain

Searching deeply - own domain
- Exploring outside domain
- Assessing hypothesis
- Exploring - own domain

Exploring outside domain
- Assessing hypothesis
- Exploring - own domain

Assessing hypothesis
- Exploring - own domain

Exploring - own domain

Categories with Importance Rankings
Where’s the impact, the deterrents?

- Greatest advancements associated with visualization of data
- Knowledge of brain anatomy (people, information resources and tools) has pivotal role in moving research forward
- Difficulty locating specifics on protocols, instrumentation, measurement, and results
- Retrospective, non-digital literature often ignored
- Background work outside specialization and assessing findings and project risks require most time and effort
Practices and implications

• Routine, high reliance on PubMed

  Arrowsmith literature mining could be widely used if well integrated with current resource. However, this raises questions about how much relevant information is being missed and how to best integrate complementary databases.

• Review articles are essential for keeping up with information and for learning in new areas.

  Need more systematic field-wide and interdisciplinary reviewing services.
## Weak and strong information work

Extension of Herbert Simon’s conceptualization of weak / strong methods in science (Simon, Langley, and Bradshaw, 1981) and Kuhn’s revolutionary / normal science to information search and use activities.

<table>
<thead>
<tr>
<th>Weak</th>
<th>Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ill-structured problem space</td>
<td>Structured problem space</td>
</tr>
<tr>
<td>Unsystematic steps</td>
<td>Systematic steps</td>
</tr>
<tr>
<td>Low domain knowledge</td>
<td>High domain knowledge</td>
</tr>
<tr>
<td>Data driven</td>
<td>Theory driven</td>
</tr>
<tr>
<td>Seek and search</td>
<td>Recognize and calculate</td>
</tr>
</tbody>
</table>
Weakest and strongest are highest impact

And best targets for information system and service development

• Weak information work is the most arduous and most speculative
  • background and feasibility in new areas
  • literature based discovery and hypothesis testing
  • ontology development for data repositories

• Strong information work is most routine and codified
  • instrumentation and methods fact-finding
  • management of literature and data, metadata
  • standards development
  • dissemination
What are we learning from these and other studies?

• Finding articles to read — left-to-right, top-to-bottom — is even less of an accurate representation of scientific literature use than it ever was.
  • We “read” less and less every year, yet are even more engaged with the literature.
• Researchers use sophisticated techniques to mobilize information according to varied research needs.
  • these information needs vary with discipline, research problem life cycle (both local and global), and particular research strategies, as well as with the varying affordances of current technology.
  • the pace of evolution and innovation in these techniques is increasing,
    • and increasingly driven by researchers themselves and not information specialists or publishers.
• Some of these techniques — literature-based discovery for instance — threaten radical and possibly even disruptive changes in researcher behavior and interests.
• Others, such as using domain ontologies and XML encoded data for searching, navigation, and browsing, promise more gradual, but perhaps more widespread and consequential changes.
What does this mean for STM publishers?

• It is safe to say that researchers will be trying to cope with the avalanche of information by applying digital technologies and new strategies for literature use in ways that require increasingly fine-grained computational access to information in scientific articles.

• Unfortunately current methods for studying literature use provide little understanding of what exactly researchers do, and why, and even less about what works … so it is not surprising that we can little predict the direction of innovation — or the results of our interventions.

• But, we’re working on that.


Contact information

Carole Palmer

clpalmer@uiuc.edu

Allen Renear

renear@uiuc.edu

Scientific Communication Initiative

http://sci.lis.uiuc.edu/

__________________________________________

Comments and Questions?