Strategic Reading and Scientific Discourse

Allen H. Renear[1] and Carole L. Palmer[1]
[1]Center for Informatics Research in Science and Scholarship
University of Illinois at Urbana-Champaign
{renear, palmer @illinois.edu}

Abstract
As scientific ontologies are integrated into the publishing workflow many enhancements to scientific communication will be possible, including improved support for text mining, information extraction, and literature-based discovery. One enhancement that is not so obvious, but will almost certainly have profound effects on the daily lives of working scientists, is support for the long-standing practice of *strategic reading*.

Keywords: ontologies, reading, scientific publishing

1. Introduction

The integration of scientific ontologies into the publishing workflow will bring about the long-anticipated revolution in scientific publishing, supporting not only text mining, information extraction, and literature-based discovery, but, most importantly, an intensification of the common practice of strategic reading. (This paper is based on our recent analysis in "Strategic Reading, Ontologies, and the Future of Scientific Publishing," *Science*, August 14, 2009 [1]).

2. Strategic Reading

Coping with vast amounts of complex information is a challenge faced by every scientist. To meet this challenge scientists have always used a variety of strategies to avoid unnecessary reading: indexing and citations as indicators of relevance, abstracts and literature reviews as surrogates for full papers, and social networks of colleagues and graduate students for filtering and recommending literature. When they do engage directly with the literature scientists do not typically read individual articles, but rather work with many articles simultaneously to search, filter, compare, arrange, link, annotate, and analyze fragments of content, in order to gather information as efficiently as possible. This behavior, which we call *strategic reading*, has been well-documented for both digital as well as print media [2][3].

As online indexing and navigation systems (such as PubMed, Scopus, Google Scholar, Citeseer) have become more sophisticated and widely used, strategic reading practices have intensified[4][5]. Today scientists often search and browse as if they were playing a video game. They rapidly move through resources, changing search strings, chaining references backward and citations forward. By note-taking or cutting and pasting, they extract and accumulate bits of specific information, such as findings, equations, protocols, and data. They make rapid judgments—such as assessments of relevance, impact, and quality—as they formulate and iteratively refine queries.

Based on longitudinal studies done by Carol Tenopir, Donald King, and colleagues on e-journal adoption trends, there is strong evidence that over the last decade STM users are "reading" more paper at a faster pace [7][8][9]. The total time spent reading journal articles has risen only a little, whereas the number of journal articles read per year has gone up much faster and appears to be growing still. The number of articles read (as distinguished from those merely browsed) by scientists was ~50% higher in 2005 than in the mid-1990s. Furthermore, though the average reading time per article did not change much from 1977 to the mid-1990s (48 versus 47 minutes), it started falling in the mid-1990s and is now just over 30 minutes per article. It is likely that the "browsing" time per article browsed is falling similarly, and that the ratio of articles browsed to articles read is growing.

Most current strategic reading relies on human recognition of scientific terms from plain text, perhaps assisted with simple string searching and mental calculation of ontological relationships, combined with burdensome tactics such as bookmarking, note-taking, and window arrangement. But with tools designed to take advantage of current enhancements, such as links to online databases, computationally available terminological annotations, and formal ontologies supporting automatic inferencing, it is possible to efficiently support specialized views, navigation, inferred results, and other processing.

Tools to support strategic reading based on scientific ontologies are already being used in the biomedical sciences [10]. One example is Textpresso[11], an ontology-based mining and retrieval system that works with prepared collections of articles, split into sentences and annotated with terms from 33 ontology categories, three of which correspond to the Gene Ontology ontologies. Results screens present a ranked list of sentences within a ranked list of articles, with term highlighting, and links to articles and external databases. Reading the sentences of an article in relevance order rather than narrative order is an example of strategic reading within an article. An example of strategic reading across a collection is provided by Information Hyperlinked over Proteins (iHOP)[12], which uses genes, proteins, NCBI taxonomy identifiers, and MeSH headings to create a network of sentences and abstracts for searching and navigating MEDLINE abstracts, presenting configurable pages of ranked lists of sentences retrieved from many abstracts.

2. The Coming Revolution in Scientific Publishing

The 1980s saw renewed anticipations of a coming new world of advanced scholarly communication taking advantage of computer networking and digital information [16][17]. This imagined world included advanced navigation, discipline-specific tools for browsing and analysis, searchable stand-off hypertext linking, data-driven user-modifiable diagrams, computationally available information items (such as computable equations and chemical formulae), structured annotations, and so on— an updated version of the grand old dream of radical new research functionality described by visionaries such as Paul Otlet, Vannevar Bush, Douglas Engelbart, and Ted Nelson.

Although much of this functionality had already been implemented in experimental hypertext systems [6], the predicted revolution never happened. Even now, in 2009, few of the features anticipated in the 1980s are available for general use, and those that have been realized are but pale versions of imagined and prototyped originals. It is true that there were substantial changes in scientific publishing during the 1990s: nearly all journals came to include a digital version that was distributed over the internet and linked from indexing

systems. But these changes provided very little of the functionality explored in the hypertext systems of the 1980s.

In retrospect we can see that in the 1990s every aspect of computing, from hardware to supporting social infrastructure, was inadequate for the emergence of a high-function scientific communication system. However since then there have been extraordinary improvements in the functionality, interoperability, and efficiency of basic networking, hardware, and software. Key basic standards and protocols have been developed and widely implemented; new software engineering strategies such as object oriented programming and conceptual modeling have been widely adopted; powerful new software applications have been developed and diffused; and user interfaces are considerably more effective and compelling. Strategies for distributed development and interoperable tools and data have been developed and tested.

Particularly important for the changes underway is the widespread adoption of a standard serialization language (XML) with associated standards and tools (XSLT, XQuery, XPath), providing a high level of interoperability at the data structure serialization level. Interoperability at the level of logical syntax is provided by the rapidly developing standards and technologies of the semantic web (RDF, OWL, SPARQL, SWRL). But most important is the development of scientific domain ontologies, which promise the semantic interoperability needed to realize the anticipated functionality.

Originally designed to support the sharing and integration of scientific data, these ontologies will increasingly be integrated into the scientific publishing workflow. Once they are deployed as part of digital scientific literature these ontologies will of course enhance text mining, information extraction, and literature-based discovery—but just as importantly they will transform how scientists "read" the narrative prose of scientific literature.

4. Finishing the Job

Some practical challenges remain.

For ontology-aware reading tools to function well without preprocessing, terminological annotations must be included in, or mapped to, the XML encoding of articles during the publishing production process, connecting names and phrases in narrative text with standard terminology. The recent convergence within the STM publishing community on a single XML schema for the representation of scientific articles: the National Library of Medicine Journal Archiving and Interchange Tag Suite provides a promising shared context for recording terminological annotation, but specific strategies that are economically sustainable within the current context of STM publishing workflows need to be developed,

In addition, "legacy data," the articles already published and stored in repositories, must be accommodated and retrofitted with terminological annotation.

Finally, to seamlessly relate terms within evolving ontologies that clarify meaning and guide inferencing, and as well as to connect terms with relevant databases, systems making use of "service-oriented architectures", that support interoperable communication over the network, will be required.

Further discussion of what changes in scientific publishing are likely to take place, and realize "semantic publishing", can be found in Shotton[13].

5. Research Needed to Align Tools and Practices

A challenge of a different sort is the limited empirical research on how scientists using digital resources read and engage with texts in the course of research. Many of the traditional approaches to evaluating information systems, such as retrieval precision and recall or satisfaction measures, do not provide the kind of analysis needed to guide the development of strategic reading technologies.

If we want to understand the fast-paced and subtle tactics, interactions, and intentions involved in using and applying the literature in online environments, then methods need to be applied that capture what scientists actually do and value as they gather, review, and manipulate texts and work with them over time. We know, for instance, that scientists often have trouble locating very problem-specific information (on methods and protocols, for instance) and that the occasional exploration of results from another discipline can have considerable impact on progress or the direction of research [14][15].

These are the kinds of information behaviors that we need to understand more fully to design tools that go beyond search and retrieval to support creative strategic reading.

6. The Persistence of Reading

Even as searching, mining, processing, and annotation become more and more sophisticated, narrative text will remain vital to scientific discourse. Scientists will not give up reading, for risk of losing the unique context and nuance provided by the flexibility and subtlety of natural language [18].

Strategic reading tools that take advantage of scientific ontologies respond directly to the entrenched value of strategic reading in the daily work of today's scientists. These tools augment the unique effectiveness of natural language narrative by bringing the power of automatic processing to bear on structured representations linked to shared knowledge bases. These new technologies will not replace reading—they will support and enhance the long-standing practice of reading strategically.

References

1. Renear, A.H., Palmer, C.L.: Strategic Reading, Ontologies, and the Future of Scientific Publishing, Science, 325, 828-832 (2009).
2. Bishop, A.P.: Document Structure and Digital Libraries: How Researchers Mobilize Information in Journal Articles. Information Processing and Management. 35, 225 (1999).
3. Schatz, B. et al.: Federated Search of Scientific Literature. Computer, 32, 51 (1999).
4. Nicholas, D., Huntington, P., Williams, P., Dobrowolski, T.: Re-appraising Information Seeking Behaviour in a Digital Environment: Bouncers, Checkers, Returnees and the Like. Journal of Documentation, 60, 24-43 2004).
5. Nicholas, D., Huntington, P., Jamali, H. R., & Dobrowolski, T.: Characterising and Evaluating Information Seeking Behaviour in a Digital Environment: Spotlight on the 'Bouncer'. Information Processing and Management, 43, 1085-1102 (2006).
6. Conklin, J.: Hypertext: an Introduction and Survey. Computer, 20 (1987) 17-41. Expanded version: Conklin, J.: A survey of hypertext (MCC Technical Report STP-356-86, Rev. 1). MCC Software Technology Program, Austin, TX, (1987).
7. Tenopir, C., King, D.W.: Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns. D-Lib Magazine, 14, 11/12 (2008).
8. Boyce, P., King, D. W. , Montgomery, C., Tenopir, C.: How Electronic Journals are Changing Scholarly Reading Patterns of Use. The Serials Librarian, 46, 121-141 (2004).
9. King, D. W., Tenopir, C., Montgomery, C., Aerni, S. E.: Patterns of Journal Use by Faculty at Three Diverse Universities, D-Lib Magazine, 9, 10 (2003).
10. Kim, J., Rebholz-Schulmann, D.: Categorization of Services for Seeking Information in Biomedical Literature: A Typography for Improvement of Practice. Briefings in Bioinformatics, 9, 452-465 (2008).
11. Muller H.M., Kenny E.E., Sternberg P.W.: Textpresso: an Ontology-based Information Retrieval and Extraction System for Biological Literature, PLoS Biology, 2, 11 (2004).
12. Hoffmann, R., Valencia, A.: A Gene Network for Navigating the Literature. Nature Genetics, 36, 664 (2004).
13. Shotton, D.: Semantic Publishing: The Coming Revolution in Scientific Journal Publishing. Learned Publishing 22, 85-94 (2009).
14. Palmer, C. L. Adapting digital information to scientific practices. STM Spring Conference, Cambridge, MA, 24-26 April. International Association of Scientific, Technical & Medical Publishers, (2007).
15. Palmer, C. L. Weak Information Work in Scientific Discovery. Information Processing and Management, 43, 808-820 (2007).
16. Yankelovich, N., Meyrowitz, N.K., van Dam, A.: Reading and Writing the Electronic Book. IEEE Computer, 18, 16-30 (1985).
17. Coombs, J.H., Renear, A.H. DeRose, S.J.: Markup Systems and the Future of Scholarly Text Processing. Communications of the Association for Computing Machinery (CACM), 30, 933-947 (1987).
18. Blake, J.A., Bult, C.J.: Beyond the Data Deluge: Data Integration and Bio-Ontologies. Journal of Biomedical Informatics, 39, 314-320 (2006).