(Bee)Dancing on the Boundary Between PIM and GIM

Shannon Bradshaw
Department of Mathematics and Computer Science
Drew University
sbradsha@drew.edu

Marc Light
Department of Linguistics and SLIS
The University of Iowa
marc-light@uiowa.edu

David Eichmann
SLIS
The University of Iowa
david-eichmann@uiowa.edu

1. PROBLEM AND POSITION

In the life sciences, 40,000 new articles are published each month, adding a large number of findings to an already immense area of knowledge [15]. Literature review is difficult, increasing the chance of duplicated effort. It is also time-consuming, reducing the amount of resources scientists have to pursue new discoveries. One approach to this problem would be to fund a talented team of researchers to curate all published knowledge in a structured database.

However, it is our position that the information gathering activity of individual investigators provides both an effective means of self-curation of published knowledge and an effective vehicle for communicating this information to researchers who will need it. Stated more concretely, our position is that researchers need an information system that enables them to:

- Search a variety of sources as usual, but with the ability to read and annotate (e.g., highlight) electronic documents as easily as they do with paper.
- Organize documents with annotations, lab notes, data, etc. and view the assembled information through a variety of lenses.
- Share these "artifacts" of information gathering with others in their research group and larger community.

For the individual, this type of system will reduce effort by integrating search, reading, annotation, and organization. It will lessen the need for piles of hard copy that are difficult to maintain and cumbersome to carry around and help avoid redundant search for something once found and later lost.

In addition, such a tool will enable distributed, grassroots curation of knowledge in the sciences through normal literature search activity (one flavor of Erickson's Group Information Management (GIM) notion [8]). By bundling together the information they gather and annotations they create into packages we call "knowledge artifacts" (KAs)[22], investigators create summaries of a space of related research that are not unlike the material from which one constructs a review article. While we do not claim that the information assembled in a KA is as easy to digest as a review article, such information will save time for researchers whose information needs touch on the same set of documents. In addition, a KA allows us to present this information to the investigator in the context of the literature itself, and minimize the degree to which users must trust the system to be accurate (in contrast with structured databases such as SwissProt).

2. A (REAL) MOTIVATING EXAMPLE

In a recent genomic analysis of a red tide dinoflagellate species, a graduate student in the Bhattacharya Lab at the University of Iowa found a gene encoding histone H2A.X. He was unaware of the specific function of H2A.X or what had previously been reported about histones in related organisms. A time consuming web search revealed the known function of H2A.X in related organisms and that H2A.X had never been reported in dinoflagellates.

Using the type of system we propose, this graduate student would digitally organize the material used in answering his questions about histone H2A.X. He would use the system to search the Web and intranet resources. He would then assemble documents of interest and highlight important passages, add margin notes, and otherwise annotate on top of HTML and PDF documents. He might also generate lab notes summarizing what he has learned and include these notes with the annotated documents. Finally, the researcher would save all this information in a knowledge artifact. The researcher might then view the KA through a variety of lenses, looking at, for example, all annotated passages containing a particular set of words such as those indicating H2A.X's function in an organism, thereby enhancing his ability to analyze the literature and provide similar benefits for others with whom the KA is shared.

Others may then reuse the fruits of these efforts. Any KA generated for a gene in this example, may be reviewed by the same or other researchers in other work on that gene. In an information gathering task such as this it is unlikely the graduate student will remember all the genes for which he constructed KAs. Even if he did, another researcher in his lab probably would not know this and would likely not think to ask. The approach we propose solves this problem by enabling the graduate student to post his KA to a repository shared by other researchers of his choosing. Another researcher may search her community's archive of KAs for information about histone H2A.X. Alternatively, as she is searching PubMed[1] for information about H2A.X, her sys-

[1] http://www.ncbi.nlm.nih.gov/entrez

tem may employ a just-in-time[2] approach [5] to automatically poll the KA archive. Using her actions (searches, documents viewed, passages highlighted, etc.) as a means of matching her to useful information, the system would find the KA on H2A.X and make it available to her.

KAs provide a means not only by which a researcher can manage his or her personal archive, but also a means where a group of researchers may collaboratively review the literature on a specific gene or other topic and make this information available to the wider bioscience community. This strategy uses the questions scientists are asking (or have asked at least once) to identify the information likely to be of interest to the rest of the community. Furthermore, our approach integrates both the capture and subsequent communication of this information with the most common means of gathering information in the biological sciences, that being literature search. Finally, as many studies demonstrate, readers prefer to view information providing answers to their questions in the context from which they were extracted. For the problem we are attacking, our approach shows the reader not simply the sentence, paragraph, and document from which a single fact was extracted but also the space of related research to which that fact contributes something new.

3. IS THIS REALLY A PROBLEM?

The total time scientists spend in information gathering and knowledge acquisition is significant. Publisher and industry surveys of scientists in university and other settings indicate that researchers, spend an average of 15-35% of their time searching, gathering, and reading information necessary to their work [18, 14]. Each year, scientists spend 166 hours discussing information, 290 hours reading, and 98 hours preparing presentations [20, 1]. We hypothesize that scientists would have greater time to focus on new work if maintaining and reviewing what they have read once were easier and if information assembled by others could be more easily consumed.

With several large knowledge curation projects underway in biology, it is reasonable to ask if these resources take the place of reading research articles. The abundance of other forms of web-based information makes this question even more important. Two prominent curation projects are SwissProt and Mouse Genome Informatics (MGI). These efforts focus on problems such as identifying the functions of genes and proteins, extracting semantic relations (e.g., protein interaction), summarizing the basic biology of a gene, etc. Even with fairly detailed information on such specific question available, the literature itself remains a primary source for assembling the current state of human knowledge.

Scientists in general report that journal articles are considered more important than other sources of information [1]. Such findings are echoed in our own study of information needs in the biological sciences. We together with several other researchers collected data for a study of the information needs of biologists in a variety of specialties as part of the Genomics Track for TREC 2004 [16].

We collected descriptions of 75 information needs from 43 subjects at 22 organizations in the United States and Britain. We asked each subject several questions regarding recent information they had encountered. For 69 of 74 (93.2%) responses to the question, "Where would you usually look to satisfy this information need?", the subject reported using PubMed Entrez, MEDLINE, or some other source for finding relevant literature. In addition to PubMed, 27 of 43 (62.79%) subjects reported using a web search engine such as Google or other web-based means of finding relevant research. In contrast, for only 19 of 74 (25.7%) responses to the same question did subjects report using a curated database such as SwissProt.

Further evidence of the focus on the literature was found in the frequency with which subjects reported using PubMed Entrez and other tools. 18 of 41 (44%) subjects responding to the question, "How often do you have this kind of information need?" reported that such needs occurred at least once a week. 12 (29.2%) reported going to the literature multiple times per week. Many subjects described more than one information need they have on a weekly basis.

To summarize, the results of this survey suggest that biologists rely heavily on the literature to satisfy information needs in the course of their research. Furthermore, they regularly search the literature, pointing to an important role for information systems that aid in the literature search process in bioscience research.

4. DO RESEARCHERS ANNOTATE?

Researchers spend a significant percentage of their work day reading and condensing information [23]. Many studies demonstrate that annotation is an important part of these tasks [9, 13, 10]. These studies and others found that researchers spend as much as 26% of their reading and writing time annotating on top of existing documents by highlighting, underlining, etc. They also indicate that annotation is pervasive in research-oriented reading and writing activities. Though none of these studies explore the annotation practices of biologists specifically, they do survey a wide spectrum of people engaged in research tasks including computer scientists, medical professionals, law students, and financial analysts. Each study found similar annotation behavior regardless of the group studied. We hypothesize that biologists' annotation behavior is in line with the findings above.

5. ELECTRONIC ANNOTATION?

Virtually every study on the subject agrees that scientists perform a majority of their literature search activities on-line [1, 15, 20, 24]. Scientists primarily access articles through electronic journal subscriptions. A recent study indicates that over 70% of readings are accessed electronically by faculty and graduate students [24]. Nearly all newly published articles are available in an electronic form and over 80% of articles read by scientists are those published within the last 3 years [20].

A large percentage of scientists not only access articles electronically but also read them in their electronic form. Charting studies of usage of digital libraries, the percentage of users who do their reading online has grown from 25% of users [27] in the mid 1990s to 33% [4] in the late 1990s to nearly 50% of users in recent years [6]. PhD students are far more likely to read articles in their electronic form than are faculty at research institutions. In our surveys of biologists at Iowa we found similar behavior. It appears that the

[2]A just-in-time system works in the background automatically gathering information that will be useful in the context of what the user is currently doing. Upon request, it provides this information to the user.

percentage of information read on-line rather than in some printed form is increasing with time. *But will researchers annotate on-line?*

Work on electronic annotation systems that manage to stay out of the way of the reading process has demonstrated that people will use such systems and do find them useful as research tools [10]. Readers want to be able to annotate without the need to interrupt the reading process to find the highlighter on a toolbar [23, 13, 26]. Readers become frustrated when they cannot scroll through a page without interrupting the reading process. Readers often move through a document in a way that is non-linear, jumping from one portion of a paper to another and back again [23, 10, 17, 19]. Building on a decade of work in human computer interaction in electronic reading and annotation systems [23, 10, 17, 19], we must take care to provide the affordances necessary for gathering information from research articles on-line.

6. FROM PIM TO GIM

In most fields of research, a small percentage of published articles receive most of the attention. King and Tenopir found that only 13% of articles published in Psychology journals are read by the community [20]. Within that small fraction of psychology articles, each is read an average of 800 times. Economics journal articles are read an average of over 1,200 times [11]. Cancer researchers report that articles published in the Journal of the National Cancer Institute are read by an average of 1,800 investigators [20]. In 2001, maintainers of the the LANL preprint archive report that each article was downloaded an average of 300 times per year [1]. Based on these studies, King and Tenopir estimate that the average number of readings of U.S. published scientific articles is about 900 per article.

It is extremely unlikely that each reader views an article for a reasons that are entirely independent of all others. In previous work, Bradshaw explored what people learned from a paper based on the way they cited it [2], finding that people tend to cite the same paper for the same reason. It is this type of overlap in information need that we seek to support with this work. In evolutionary biology, examples include a body of documents describing specific molecular pathways, genes, and many other topics, information that can be used to pursue a variety of research paths.

7. WILL RESEARCHERS SHARE?

KAs enable a variety of instructional as well as research uses. A PI might pass along a bundled literature review to help a new graduate student come up to speed on the project. An instructor may use the same means to make a collection of articles more accessible to a group of students taking a seminar class. An investigator may pass along her literature review to collaborators with whom she is engaged in a research project. A team of researchers may use a KA as a talking point as they do background research to develop a grant proposal. Many years of research in knowledge management indicate, perhaps surprisingly, that even small work groups (5 to 10 people) fail to share information and knowledge effectively [7]. Many times members duplicate information gathering tasks and exhibit other inefficiencies that reduce productivity. Finally, the members of many laboratories, departments, and organizations will have few inter-organizational competitive concerns, but several incentives to share. Accumulated KAs will serve as the organizational memory for each laboratory or other research group for which our approach is used. Any piece of knowledge once acquired by one researcher in the team can be acquired by another researcher in that same team with a fraction of the effort of the first investigator. Furthermore, as a laboratory evolves with the graduation of graduate students and continuous cycle of postdoctoral researchers, information will not be lost to the degree it would without such a repository in place.

Recently, there has been increased attention focused on information sharing among the members of a broader research community. For example, one community entitled *African Lakes Limnology* on CiteULike (www.citeulike.org) describes itself as relating to "Biology of the East African Rift Great Lakes: bacterioplankton, phytoplankton, zooplankton, microbial food web, etc." The archive for this community contains over 900 documents carefully categorized by hand, with summary notes in some cases. Many other active communities in the biological sciences may also be found on CiteULike. Additional precedent for bioscience information sharing is found in the NeuroScholar system [12], funded by the National Library of Medicine. NeuroScholar uses a grassroots curation approach to enable neuroscience researchers to curate a centralized database of neuroscience facts and tie it to relevant literature. Key differences between our approach and that of NeuroScholar are that NeuroScholar is not so contextualized within the literature itself and it requires a significant amount of work from participants beyond their ordinary literature search activity.

8. WILL KAs BE FOUND USEFUL?

There is a growing body of literature to suggest that people engaged in reading the same document overlap in which passages they consider useful. Most of this work looks at the annotation behavior of readers other than scientists (e.g., undergraduate students) [21]. Other studies showed that people find the annotations of others useful [26, 21]. Our own studies of reading and annotation behavior of bioscience researchers indicate that there exists a high degree of consensus as to which passages are important in a paper and that these passages account for only approximately one-third of the total text in a paper [3]. With the consensus of previous readers overlayed as highlighted passages or some other marking, new readers of a paper might skim to greater effect and in less time. In addition, this interaction mechanism can serve as a useful means of introducing the commentary of previous readers and of sparking dialog between researchers.

A KA will provide a place to organize and save everything that an individual or research group knows about some topic (e.g., histone H2A.X). It will provide a point of dialog around which collaborators can refine and extend their understanding. Two challenges will be in enabling editing and extension of KAs and in helping researchers update KAs as time goes on. A versioning system similar to that of the wiki technologies will likely be necessary [25] to enable smooth updating and rollback to previous versions. Automated queries based on vector-space representations of KAs will provide consumers of KAs with accurate pointers to new work published since the KA was created [5].

9. HELPING RESEARCHERS SCALE UP

Sharing KAs will enable researchers to leverage the work done by others. However, often a researcher will find herself in an area where she has no access to a relevant KA and where there are far too many articles to read and annotate one by one. It is our position that current text analysis tools, developed by the natural language processing research community, can be used to increase coverage.

We envision a personal system where the researcher can dump in 55 PDFs and then ask the system to extract all organism names and create an index based on these names. In the H2A.X histone example above, our graduate student would have been able to browse a list of organism names from all 55 H2A.X articles. Clicking on an organism name would take him to a list of passages containing mention of the that organism. Similarly, a user might highlight a small number of passages and then ask the client to highlight similar passages throughout an article or set of articles. Finally, a user might ask the system to extract a specific type of tabular data such as protein interactions and then edit the results, weeding out false positives.

10. BEEDANCE

We are implementing a PIM/GIM system called Beedance. Beedance serves both as a tool for managing information and knowledge in the sciences and as an environment in which we study information sharing in research and learning contexts. The name is borrowed from the waggle dance that scout honeybees use to communicate the location of nectar to their hive-mates. The Beedance system is web-based and user-managed following a model similar to citeulike.org. Users create and evolve communities of people with whom they wish to share the results of their information gathering efforts. Using the Beedance client, researchers search and browse the Web, read PDF articles, and annotate what they read with electronic highlighting, margin notes, etc. The client also allows users to build KAs by organizing web pages, research articles, images, annotations, and other information around a research question. KAs identify the path through a body of literature which provides answers to this question (or nectar) and thus the bee waggle dance analogy. The system we envision is not unlike a recommender of sorts for research articles and passages within those articles. Though the actual interaction may differ our objective is to provide researchers with a tool that can provide services such as "Others who read these three articles, also found the following fifteen articles useful." or "Others who read this article found the highlighted passages especially important."

11. REFERENCES

[1] B. Bauldock, W. Wicks, and J. O'Neill, editors. *MetaDiversity II: Assessing the Information Needs of the Biodiversity Community*, 2001.

[2] S. Bradshaw. Reference directed indexing: Redeeming relevance for subject search in citation indexes. In *Proc. of ECDL*, 2003.

[3] S. Bradshaw, M. Light, and D. Scott. Annotation consensus: Implications for reducing the burden of literature review. Submitted to BMC Bioinformatics.

[4] C. M. Brown. Information seeking behavior of scientists in the electronic information age: Astronomers, chemists, mathematicians, and physicists. *JASIS*, 50(10):929–943, 1999.

[5] J. Budzik and K. J. Hammond. User interactions with everyday applications as context for just-in-time information access. In *Proc. of IUI*, 2000.

[6] J. M. Cherry and W. M. Duff. Studying digital library users over time: A follow-up survey of early canadiana online. *Information Research*, 7(2), 2002.

[7] J. Cummings. Work groups, structural diversity and knowledge sharing in a global organization. *Management Science*, 50(3), 2004.

[8] T. Erickson. From pim to gim: Personal information management in group contexts. *CACM*, 49(1), 2006.

[9] A. Adler et al. A diary study of work-related reading: Design implications for digital reading devices. In *Proc. of CHI*, 1998.

[10] C. Marshall et al. Introducing a digital library reading appliance into a reading group. In *Proceedings of Digital Libraries 1999*, 1999.

[11] F. Machlup et al. Information through the printed word: The dissemination of scholarly, scientific, and intellectual knowledge. *Journals*, 2, 1978.

[12] G. Burns et al. Tools and approaches for the construction of knowledge models from the neuroscientific literature. *Neuroinformatics*, 1(1):81–110, 2003.

[13] K. O'Hara et al. Student readers' use of library documents: Implications for library technologies. In *Proc. of CHI*, 1998.

[14] S. Feldman. The high cost of not finding information. *KM World*, 13(3), 2004.

[15] P. Fontelo, M. Ackerman, and G. Kim. Development of evidence-based medicine resources: Bridging clinical research to medical practice. In *Proc. of the 37th Hawaii Int. Conf. on System Sciences*, 2004.

[16] W. R. Hersh and R. T. Bhupatiraju. Trec 2004 genomics track overview. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, 2004.

[17] K. Horbaek and E. Frokjaer. Reading patterns and usability in visualizations of electronic documents. *ACM Trans. on Computer-Human Interaction*, 10(2):119–149, 2003.

[18] Outsell Inc. Content user profile: Update on scientists. *InfoAboutInfo Briefing*, 7, April 30 2004.

[19] W. C. Janssen. Readup: A widget for reading. Technical Report TR-05-3, PARC, 2005.

[20] D. W. King and C. Tenopir. *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers*. 2000.

[21] C. Marshall. Annotation: from paper books to the digital library. In *Proc. of Digital Libraries*, 1997.

[22] B. Newman. *The Knowledge Management Handbook*. Springer-Verlag, 2002.

[23] K. O'Hara and A. Sellen. A comparison of reading paper and on-line documents. In *Proc. of CHI*, 1997.

[24] C. Tenopir. *Use and Users of Electronic Library Resources*. Washington, D.C., 2003.

[25] Wiki front page. http://c2.com/cgi/wiki.

[26] J. L. Wolfe. Effects of annotations on student readers and writers. In *Proc. of Digital Libraries*, 2000.

[27] H. Woodward. Cafe jus: Commercial and free electronic journals user study, 1997.