

Finding and Using Journal Article Components: Impacts of Disaggregation on Teaching and Research Practice

Robert J. Sandusky

Carol Tenopir

School of Information Sciences, University of Tennessee, Knoxville

Abstract

This paper reports the results of a study into the use of discrete journal article components, particularly tables and figures extracted from published scientific journal articles, and their application to teaching and research. Sixty participants were introduced to and asked to perform searches in a journal article component prototype that presents individual tables and figures as the items returned in the search results set. Multiple methods, including questionnaires, observations, and structured diaries, were used to collect data. The results are analyzed in the context of previous studies on the use of scientific journal articles and in terms of research on scientists' use of specific journal article components to find information, assess its relevance, read, interpret, and disaggregate the information found, and reaggregate components into new forms of information. Results indicate that scientists believe searching for journal article components has value in terms of (1) higher precision result sets, (2) better match between the granularity of the prototype's index and the granularity of the information sought for particular tasks, and (3) fit between journal article component searching and the established teaching and research practices of scientists.

Introduction

Scientists use journal articles in their work for many reasons, including research, teaching, and current awareness. The average number of articles read per scientist each year has steadily increased over the last three decades, reflecting the concomitant rise in the number of articles published and the pressure to read more to remain current (Tenopir & King, 2000; 2004). At the same time, the average amount of time spent per reading has decreased from 48 minutes per article in 1977 to 34 minutes in 2005 (Tenopir, Nicholas, & Baker, 2006). These trends – reading more articles and devoting less time to each article – are likely to continue as the both the amount of scientific information increases (National Science Board, 2006) and more of it is readily accessible through networked indexing and abstracting systems and the open Web. Scientists need new ways to locate the best literature in their field efficiently and to identify the most relevant parts of the articles quickly.

Although a majority of the articles read by scientists are now from electronic journals, and almost all use of indexing and abstracting services is from electronic versions of these services, the main methods of locating relevant information remain browsing recently-published journal issues, searching index terms that are assigned to surrogates that represent an entire article, or free text searching of the article text (Friedlander, 2002; King et al, 2003). To locate relevant articles, scientists must scan an ever increasing number of abstracts,

journals, or article texts. To locate relevant tables and figures contained with articles, scientists must typically retrieve article surrogates from an indexing and abstracting database, assess the relevance of the surrogate, and then locate the full-text of the article, even when just a table or figure would satisfy the searcher's information need. While some full text systems provide free text searching of figure or table *captions*, they do not index each table or figure, or allow retrieval of figures and tables as discrete, independent components.

CSA¹ developed a prototype system (referred to here as the *component prototype*) that provides detailed indexing of individual components of scientific journal articles – specifically, the tables and figures within the articles – to provide more efficient access to the expanding journal literature. Each table and figure from each article is treated as a first-class object in this system, in contrast to existing indexing and abstracting services, which treat complete articles or article-level surrogates as first-class objects. The component prototype thus provides access to the scientific literature at a new, finer level of *granularity*. The support of end-user search and retrieval of individual journal article components represents a move by commercial indexing and abstracting systems toward a more complete realization of what has been labeled the *advanced system phase* of electronic journal systems (Tenopir et al, 2003).

This article reports on an evaluation of the component prototype. The evaluation was designed to assess the impact on scientific work of providing a new, finer-grained means of accessing scientific journal literature, at the level of individual tables and figures drawn from scholarly journal articles. This article also addresses the following narrower issues raised by the creation of the prototype: What are the current experiences of scientists using journal articles and the indexing and abstracting systems built to support search and retrieval, and what challenges do they currently face in this environment? How are scientists currently searching for and finding discrete information components? What are scientists' expectations for systems to support searching for and finding discrete information components? What are scientists' reactions to use of a prototype component indexing system that supports searching for and retrieval of tables and figures extracted from scientific journal articles?

A limited body of prior research, particularly in the digital library community, suggests that scientists will find value in systems that *disaggregate*² journal articles into their components (Bishop et al, 2000). This paper reports on the first significant evaluation of an indexing system or digital library providing direct access to discrete components of journal articles since the reports of Bishop and her colleagues (Bishop, 1998; Bishop, 1999; Bishop et al, 2000). The term *component* here means any logical subdivision of a scientific journal article. A partial list of the possible components of a scholarly journal article include article title, section headings and subheadings, tables, figures, captions, reference list, individual references, abstract, author assigned keywords, author names, author affiliations, author contact information, article sections and subsections, footnotes, endnotes, appendices, paragraphs, sentences, noun phrases, words, and external linked information related to the paper (e.g., datasets, additional analysis, etc). A reader may encounter other terms, such as

¹ CSA was formerly known as Cambridge Scientific Abstracts.

² *Disaggregation* can be defined as dividing an object into its constituent parts.

“information component,” “document component,” “article component,” or, more generally, “document structure,” that may be considered synonyms for *component*.

This paper builds upon Bishop’s (1998; 1999) exploratory study of similar issues, using a larger sample of users and a more advanced indexing system, the component prototype. Following Bishop, the terms *disaggregation* and *reaggregation* are used to describe the *dis-integration* of an established and taken-for-granted genre, the scholarly journal article, into discrete components and the *re-integration* of discrete components by users into new assemblages, forms, and potentially, in the longer run, new genres.

Component Prototype Description

The component prototype adds functionality beyond that typical of other contemporary abstracting and indexing systems. The prototype’s advanced search interface allows searchers to construct complex Boolean queries, limit searches to specific fields of the indexing record (e.g., author; title; statistical, geographic, and taxonomic terms), and set specific limits appropriate for the nature of the prototype index, such as limiting the search to only maps, photographs, graphs, tables, or figures; or limiting the results to open access journal articles or components from articles referencing predictive models.

The results sets returned from a search conducted in the component prototype are significantly different than the results returned from a conventional indexing system. Instead of each item in the results set representing a surrogate for a complete journal article, the component prototype presents a surrogate of an individual component – either a table or a figure – culled from a journal article represented in the component prototype’s index (Figure 1).

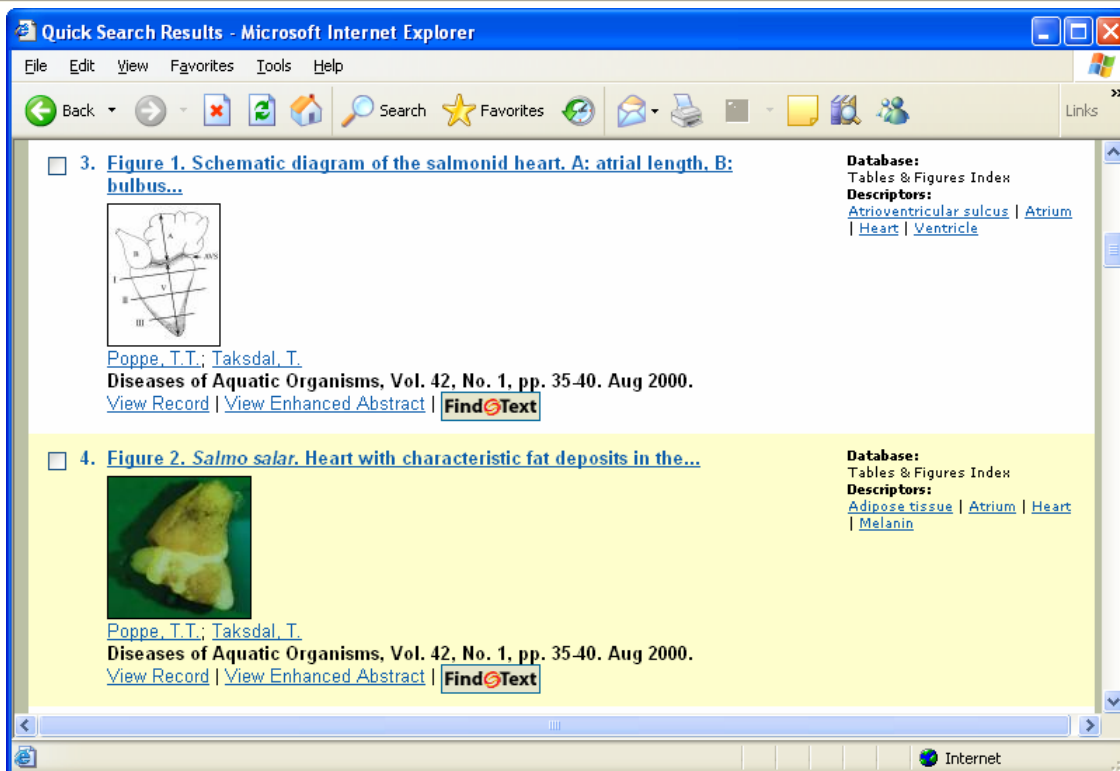


Figure 1: Display of two items in a results set returned by the component prototype developed by CSA. Each item in the results set includes a truncated caption, a thumbnail of the figure or table, citation information (author, journal name, volume, issue, page numbers), and links to the complete component surrogate, the containing article’s enhanced abstract (see Figure 3), a link to the article full text, and the descriptors assigned to the table or figure’s surrogate record. (Note: At the time the component prototype was evaluated, the article title was not displayed as part of the results set due to a design oversight.)

Selecting the thumbnail image, the truncated caption, or the “view record” link provided on the search results display takes the user to the full surrogate record for the individual component, which in most cases includes a higher resolution image of the component (Figure 2). The component surrogate display provides the full caption, indexing terms (descriptors) associated with the selected article component, including geographic, taxonomic, and statistical terms as appropriate; and the category of component. Categories of statistical terms and component types are drawn from controlled vocabularies. Examples of categories for images and maps are *Chemical Structure*, *Topographic Map*, or *Gene/Protein - Maps & Sequences*. Examples of statistical terms are *Tukey Procedure*, *T-test*, and *Polychotomous Logistic Regression*. Figure 2 shows that this figure is categorized as a “schematic” and that geographic and taxonomic descriptors, but no statistical descriptors, have been applied.

http://ca2.csa.com - View Record - Microsoft Internet Explorer

File Edit View Favorites Tools Help Links >>

III

Caption Figure 1. Schematic diagram of the salmonid heart. A: atrial length, B: bulbus arteriosus, V: ventricular length, AVS: atrioventricular sulcus. Fat deposits typically occur in shaded areas. The locations of histological transverse sections are marked I, II and III

Category Schematic

Title Ventricular hypoplasia in farmed Atlantic salmon *Salmo salar*

Author [Poppe, T.T.](#); [Taksdal, T.](#)

Descriptors Atrioventricular sulcus Atrium Heart Ventricle

Geographic Terms: Norway

Taxonomic Terms: Salmonidae (Salmonids) *Salmo salar*

New Search Using Marked Terms: Use AND to narrow Use OR to broaden

Source Diseases of Aquatic Organisms, Vol. 42, No. 1, pp. 35-40. Aug 2000.

ISSN 0177-5103

Accession Number 301-0000003097

< [Previous](#) | [Next](#) >

© 2006 CSA | [Privacy Policy](#) | [Terms and Conditions Governing Use](#) | [Feedback](#) Interface English

Internet

Figure 2: Display of a portion of the component's complete surrogate record in the component prototype developed by CSA. The figure is classified as a schematic. Note that along with a higher resolution image of the component, the full caption, general descriptors assigned to the component (Atrioventricular sulcus, Atrium, Heart, Ventricle), as well as the geographic and taxonomic terms are displayed.

The component prototype also includes an enhanced abstract, which is a surrogate for the article from which a component is culled. The enhanced abstract is based upon the existing article-level surrogate included in the conventional article-level index. Figure 3 shows that thumbnails of each of the components included in the article have been added to the standard article surrogate, providing additional information to the scientist when assessing the relevance of the article or its discrete components.

[Logout](#) | [Quick Search](#) | [Advanced Search](#) | [Search Tools](#) | [Browse](#) | 0 Marked Records | [Search History](#) | [Alerts](#)
[Record View](#) | [Return to Results](#) | [Help & Support](#)

Abstract [Mark This Record](#) | [Update Marked List](#) | [Save, Print, Email](#) | [RefWorks](#)
[U L I C H S RESOURCE LINKER](#) | [InterLibrary Loan](#) | [Document Delivery](#)

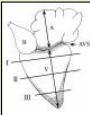
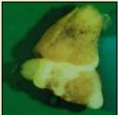
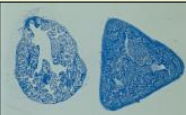
Database ASFA 1: Biological Sciences and Living Resources
Title **Ventricular hypoplasia in farmed Atlantic salmon *Salmo salar***
Author [Poppe, TT; Taksdal, T](#)
Affiliation The Norwegian School of Veterinary Science, PO Box 8146, Dep., 0033 Oslo, Norway, truve.poppe@veths.no
Source Diseases of Aquatic Organisms [Dis. Aquat. Org.], Vol. 42, no. 1, pp. 35-40, 10 Aug 2000.
ISSN 0177-5103

Descriptors Heart Cultured organisms Fish culture Fish diseases Hatcheries Temperature effects *Salmo salar* Norway

New Search Using Marked Terms: Use AND to narrow Use OR to broaden

Abstract Atlantic salmon *Salmo salar* L. parr and pre-smolts from 2 Norwegian hatcheries showed reduced weight gain, abnormal behaviour and signs of circulatory disturbances. Necropsy revealed conspicuous fat deposits around the heart to be the most consistent finding. Furthermore, the ventricle/atrium ratio was altered, with the size of the ventricle significantly smaller than normal in affected fish. Histology showed poor development or absence of the outer, compact myocardium, large numbers of fat cells and melanomacrophages in the epicardium, fibrosis, and inflammation of the compactum/spongiosum interphase. Nuclei of the inner spongy myocardium showed signs of compensatory hypertrophy. The cause(s) of this malformation is(are) unknown, but a high prevalence of other malformations in fish from the same population indicates high temperature during incubation of the eggs as a possible aetiology.

Language English
Summary Language English
Publication Year 2000
Publication Type Journal Article
Identifiers histopathology; Atlantic salmon
Environmental Regime Marine; Brackish; Freshwater
Input Center, ASFA CS0100464
Classification Q1 01346 Physiology, biochemistry, biophysics; Q3 01587 Diseases of Cultured Organisms; O 5060 Aquaculture
Update 200101
Subfile Oceanic Abstracts; ASFA Aquaculture Abstracts; ASFA 1: Biological Sciences & Living Resources
Accession Number 4793540

Tables & Figures







Figure 3: Display of the component prototype’s enhanced abstract. Note the addition of hyperlinked thumbnail images to four figures and one table at the bottom of Figure 3. Selection of one of the thumbnails causes the display of the complete surrogate for that specific table or figure (as illustrated in Figure 2).

Background

This study is grounded in the tradition of studies of scholarly communication in science, and more specifically within traditions of research into electronic scholarly publishing and electronic journal use. A complete review of the literature on electronic scholarly publishing

is beyond the scope of this paper, but reviews of that literature include Peek & Pomerantz (1998) and Kling & Callahan (2003). Within all studies of electronic scholarly publishing, this work is most closely aligned with evaluations of the effectiveness of specific systems. Rather than being a summarization of findings for an entire digital library (e.g., Borghuis et al, 1996; Bishop et al, 2000; Entlich et al, 1997; Eason et al, 2000), however, the current study focuses on a narrower topic: the impact on scientific work of indexing the tables and figures drawn from journal articles.

This study's focus on the use of journal article components is notable for its emphasis on how these scholarly materials are used in both teaching and research. Investigations of scientists' use of journal articles for purposes other than research have been rare. Hart (1998) surveyed faculty at one non-research intensive university to investigate the relationship between faculty roles of teaching, research, and service and six abstract categories of information sources, but did not investigate how faculty searched for specific forms of information; nor did he examine how they made use of the information they found. Borgman et al (2005) examined the general information seeking behavior by geography faculty in support of undergraduate teaching as part of the ongoing evaluation of the ADEPT (Alexandria Digital Earth Prototype) project. One theme among the findings on information seeking in the ADEPT project concerns information seeking for research and teaching. Borgman and her colleagues found that geographers "(a) ...found it easier to articulate their research activities than their teaching activities; (b) research and teaching activities were viewed as mutually reinforcing; (c) they continually scan their environment for information sources and glean for both purposes; and (d) they rely on their own research data as information sources for teaching." (p.648) The current study had use of figures and tables extracted from scholarly journal articles in both research and teaching as its primary focus.

Several studies have examined how scientists read and scan journal articles in order to identify relevant items for closer reading and use in research, writing articles, etc. Olsen (1994) noted the disciplinary differences regarding which journal article components are most useful for making article relevance assessments. Abstracts were found to be important for chemists and sociologists, while humanists relied upon the first few paragraphs of an article. Chemists also relied upon figures and figure captions, confirmed by findings from the CORE project (Entlich et al, 1997) and established for a wider range of disciplines by Bishop (1995; 1999). Stewart (1996) interviewed thirty-nine users of the CORE project about the importance of ten features of electronic journal systems. Seventy-three percent of her interviewees indicated that "browsing graphics to determine the value of an article" was a "very important" feature; another 15% indicated it was "important" and only 13% indicated that this feature was "not important." Figures and tables usually represent the central ideas and concerns of a scientific paper and can serve as either a synopsis or summary of an entire article, or as a means of assessing article relevance and a user's decision of whether to copy, print, or more fully read the article. Simpson (1988), Olsen (1994), and Bishop et al (2000) also identified the important relationship between the user's purpose or task and the approach the user takes to scanning or reading an article or its surrogate.

The DeLIver testbed, in operation between 1996-1998, attempted to provide end-users with the capability to retrieve articles by exposing information culled from specific components

within complete documents. This capability was derived from customized indexing and retrieval and display mechanisms that were applied to SGML files provided by the publishers of more than 50 physics, engineering, and computer science journals. DeLIver was able to support innovative search capabilities including specific searching of section headings, table and figure captions, table text, and cited references (Bishop, 1998; Bishop, 1999; Bishop et al, 2000).

Bishop examined how scientists used article components in support of research in the context of the deployment of the DeLIver digital library. She sought to understand “the process of using components to identify documents, read them and incorporate them into one’s ongoing work” (1999, p.261) and develop “...a preliminary framework for understanding journal article component use among academic researchers in science and technology disciplines” (p.270). Bishop’s framework consists of five elements: (1) finding relevant documents; (2) assessing document relevance; (3) reading documents; (4) disaggregation; and (5) reaggregation. Bishop examined use of journal article components in the broad context of scientific work, foregrounding the situated nature of scientific research:

“While researchers hunt for and read articles, they are both thinking and doing something. Aside from turning pages or clicking, they are making notes on scraps of paper, annotating, sorting articles into piles. Extracted pieces of information are set aside or sent to colleagues. Information in article components is reaggregated into a range of transitory compilations and transitional documents before being integrated into the downstream document that represents the final product of the user’s own work” (Bishop, 1999) p.262.

The above description of the complexity of the act of reading and use illustrates the importance to researchers of support for associated activities such as thinking, reflection, annotation, reaggregation, and intra-document (e.g., “turning pages”) and inter-document mobility (e.g., “sorting documents into piles”) (Olsen, 1994; Stewart, 1996; Bishop, 1999; Institute for the Future, 2002b) .

Four of Bishop’s findings are relevant to the evaluation of the tables and figures index prototype and have also been noted by subsequent researchers. First, figures are likely to provide particular advantages to users, such as being more effective summaries of the content of the article compared to subject headings or other descriptors as well as being more effective cues for remembering and recalling the gist of an article’s content. Second, it is difficult to design a system that will satisfy a wide variety of users with differing work practices: there is no single, a priori correct design. Affordances built into the interface are critical as are effective signals from designers to users about innovations in this genre of systems. “[T]echnical affordances limit or promote component use.... [W]hat people did before and after their use of article components suggest that the interoperability of the technologies used at each stage of the process affect the use of components overall” (Bishop, 1999) p.271. Regarding users’ expectations from their experiences using this genre of systems, Bishop states “interface design problems also contributed to lack of use. Some people did not use these features in DeLIver because *they did not notice them or could not figure out how to use them*. However, another factor in limited use is that *people do not*

expect these capabilities and thus are not inclined to notice them, let alone to seek their availability” (Bishop, 1999) p.272 [emphasis added]. Third, Bishop found that her interviewees reported assessing the relevance of full-text articles by scanning graphics, list of references, etc., in addition to reading article titles, authors and their affiliations, and abstracts. This kind of scanning or reading of specific journal article components is frequently done in an idiosyncratic way, typically in a non-linear sequence. This practice was also noted by Olsen (1994) and in reports from the Stanford e-Journals project (Institute for the Future, 2002b). Fourth, like Stewart (1996), Bishop found that researchers have concerns about the de-contextualization of figures when they are extracted from the surrounding explanatory text in the journal article. Such decontextualization may lead to misunderstanding of the research reported in the article and then subvert the reader’s own later contributions to the literature.

The Chemistry Online Retrieval Experiment (CORE) created and user-tested a digital library of chemistry journal articles. One major component of the CORE project was the extraction of these journal article components: figures, tables, equations, and schematics (Entlich et al, 1997). CORE provided access to journal article components via text searching and retrieval of complete articles, but they also created “an interface in which articles were represented by merely a list of authors and small-size images of the pictures [i.e., thumbnails]” (p.122) CORE also attempted, unsuccessfully, to classify figures and tables “directly rather than accessing them through the articles.” Their studies of usage of CORE revealed that tables and figures were of great value to users in assessing article relevance (see also Stewart, 1996), and they concluded that “interfaces based on them [tables and figures] will be more successful” than interfaces based upon text searching alone.

Lu et al. (2006) developed a system for categorizing figures automatically extracted from PDF documents sampled from the CiteSeer digital library. Their system uses a machine learning approach combined with automatic image processing algorithms to support the classification of extracted figures into the following categories: photograph, 2-D plot, 3-D plot, diagram and others. They compared the performance of their figure classifier to human classification of the same test set of more than 5,000 figures extracted from scientific papers randomly selected from CiteSeer in order to calculate precision and recall ratios as a first cut evaluation of their algorithms. Their classification system has not yet been integrated into a full information retrieval system, and no evaluation of the efficacy of this approach for scientists has been attempted. Liu et al. (2006) describe a system that automatically extracts table metadata from PDF documents in CiteSeer. In their approach, PDF documents are converted to formatted text files, table “candidates” are detected and confirmed, and the table structure is recognized (e.g., frame, metadata, layout, content metadata).

Schlieder & Meuss (2002) argue that the increased use of structural markup, such as XML or SGML, provides an opportunity to take advantage of document structure, supporting users in their ability to create precise queries that limit the occurrence of search terms to more specific contexts. Their system uses the document tree expressed in XML, a relatively low level of granularity, instead of the components of greater semantic importance to the user, such as tables and figures.

Florance & Marchionini (1995) conducted a study of physicians using journal literature to solve patient-related clinical problems. They documented the ways physicians evaluated document surrogates, citations, and the articles in service of problem solving, noting that primary focus remained on problem solving rather than searching or documents. They concluded that “a customized document surrogate,” such as “a multi-layered hypertext biomedical article, whose surrogate provides links to defined areas within the document,” would be required for “effective literature-based clinical problem solving” (p.162). Demner-Fushman et al (2006) explored the use of automatically identified outcomes-related information in medical journal articles for support of clinical decision making. Based upon automatic parsing of the texts of both structured and non-structured abstracts, this study suggests that document surrogates could provide better support for clinical, time constrained decision making if outcome statements were a distinct component of the citation (p.59).

Taking the perspective of the journal publisher, Stanford University’s e-Journal User Study noted that “scholars do not always read articles in linear fashion from beginning to end.... In their reading practices, many scholars thus disaggregate journal article content, reordering sections of the article as they see fit” (Institute for the Future, 2002a). They conclude that supporting content disaggregation to support researchers’ information practices – particularly reading or scanning to support relevance assessment as well as reading for understanding (Institute for the Future, 2002c) – is an opportunity to improve the value of e-journal offerings. They stop short, however, of recommending support for direct retrieval of journal article components and suggest that publishers or aggregators provide an indication of the “number of tables or graphics” as part of the information available to searchers (Institute for the Future, 2002b).

Burbules & Bruce (1995) examined the emerging and potential changes on scholarly communication, writing and the academic publishing industry, such as increased hypertextuality and integration of multimedia into writing, augured by the rapid adoption of ICTs by academic researchers. “Just as a computer with a word processor is not just a fancy typewriter, and just as new technical methods do influence the style and content of writing, the capacities of electronic networks for storage, retrieval, and dissemination are altering the way scholars produce writings and their intellectual relations to one another” (p.12). Kircz (1998) suggests that, given the capabilities of digital systems, the traditional format of the journal article be abandoned. He proposes a future more radical than improved access to traditionally-structured journal articles, one in which authors must re-conceive their writing / article production processes by “break[ing] apart the linear text into independent modules, each with its own unique cognitive character” (p.217). While his proposal implies that authors must write in a hypertextually aware manner, there are also impacts for A&I services as well: classification terms will have to be associated with the modules of particular articles, not just the entire article. He later describes (Kircz, 2002) article modules as “textual, pictorial, or other representation[s] of an amount of information that in itself is sufficiently comprehensive to convey meaning for a reader” (p.29), including specific molecules, numbers, etc., as expressed through specialized markup languages such as Mathematical Markup Language (MathML) and Chemical Markup Language (CML), each of which “must carry inseparable metadata with them” (p.30) to be useful. His notion includes typed links between modules (e.g., linking from text_1 by author_A to text_2 by author_B), where the

modules may be physically distributed from each other, a return to the notion of typed, bi-directional links from classic hypertext research. Other researchers are working to develop tools to support authors creating rich semantic structures as they write scientific journal articles (De Waard, 2005).

Wilensky & Phelps (1996) describe multivalent documents, which “‘slice’ a document into layers of more uniform content, to which additional layers may be added at a later time at equal status” (Wilensky & Phelps, 1996) (p.100). Documents are decomposed from monolithic, complex single objects into multiple constituent objects. Each constituent object has a number of specific behaviors associated with that object. Interesting examples of the use and value of multivalent documents include the capability to identify and manipulate data contained in tables, such as resorting table rows; support of user annotations and extensions to existing documents, associated in the main document’s document tree as embedded documents; and lenses to control a geometric, movable, resizable portion of the document to apply various effects to the display (Phelps & Wilensky, 2001) (p.62). Their model is a radical re-conceptualization of the nature of the digital document itself, as layers plus behaviors, rather than the evolutionary steps taken in the DeLIver system or the purely structural re-conceptualization proposed by Kircz.

Methods

Multiple methods were employed to collect data for this study.

- 1) Pre- and post-search questionnaires (to measure characteristics of the participants, prior knowledge, experiences with, and potential uses of journal article component indexing)
- 2) Observations of participants while they searched on topics provided by the researchers and at least one topic of the participant’s choice (to measure potential utility and uses, detect usability problems, and identify functional requirements), and
- 3) Structured diaries kept by participants as they searched the test database on their own for topics of interest to them (capture query terms and modifications; outcomes of individual searches)

One of three researchers traveled to one of the nine research sites to introduce the system and its capabilities, introduce the research plan, and conduct observations of scientists’ use of the journal article component indexing prototype. The on-site sessions consisted of introductions and an overview of the research project and procedures; distribution and collection of informed consent forms; distribution and collection of the pre- questionnaire; an introduction to the journal article component indexing approach and the prototype; live demonstration of the journal article component indexing system followed by a short practice session with individuals’ usernames and passwords; and wrap-up, including distribution of the structured diaries and finalization of times and locations for the individual observations. The three research team members conducting the site visits used the same materials in order to minimize the differences in how the participants were introduced to the project.

There are some obvious disadvantages to this process. By providing the participants with an introduction to and hands-on practice with the system features under evaluation, some bias is introduced. It may be expected that the participants would have a heightened awareness of the new system features compared to their awareness of the same features in a more natural research setting. However, given the time constraints and the limited nature of the research questions posed, this tradeoff was deemed acceptable. This study is not designed to be a situated, naturalistic study of the use of this system in the life of scientists; nor is designed to be a comprehensive study of the information practices of scientists in a wider context.

During the introductory session participants were asked to complete the pre-test questionnaire that provided us with demographic information, plus information about their use of journal article databases, journal article components, and anticipated use of a journal article component indexing system. At the end of the study, after participants turned in their structured diaries, a post- instrument was distributed electronically. The post- questionnaire contained many of the same questions as the pre-questionnaire to support before and after comparisons. The collection of information from the pre-test/post-test method will allow us to measure how exposure to the journal article component prototype influenced their perceptions of the utility of journal article component indexing in their work.

During the introductory onsite session, participants were asked to volunteer to return for a one-on-one observation session. Most agreed to do so. The observations were conducted based upon the recommendations of Monk et al (1993) for conducting an “obtrusive” observation of user behavior. In this style of observation, participants are provided with a short set of “typical” system tasks plus some open-ended tasks and asked to “think aloud” as they use the system. The system tasks were followed by a short debriefing in order to ascertain participants’ impressions of the journal article component prototype and allow participants to offer suggestions for improvement and express how the deep indexing features might or might not benefit their work in the future.

Structured diaries (Beheshti, 1989; Hyldegård, 2006) were used to probe more deeply into how journal article component indexing might be useful to researchers. Each participant was asked to conduct a minimum of five and up to ten searches on the prototype during the two weeks following the site visit. They could search on any topic of interest to them, using any of the features available on the system. Each participant received a unique password and unlimited use of the system. For each search they filled out a set series of questions plus open-ended reactions to the search process or results. The use of such diaries offers the best mix of structured questions, open-ended responses, and individual leeway in the topics searched. It provides us with both quantitative and qualitative data on real uses of the database and system features in a more natural setting than the observations provided. Participants have time to search and reflect on their experience without the pressures of an observer or without the somewhat artificial nature of the observed session. On the negative side, the amount of open-ended information that each participant recorded in his or her diary varied.

Together, these multiple methods provide us with a variety of information on the potential uses of journal article component indexing and problems with the current prototype. Using

multiple methods increases the validity of the measurements. Different aspects of the research problem—including current searching behavior, likely uses of figures and tables, and reactions to the system features—are covered by more than one type of question or one method. In this way we increase the likelihood that we are measuring all aspects of the real issues. Reliability, or the likelihood that the questions we are asking will be accurately and consistently answered by respondents, is also increased by measuring the same things in more than one way. In addition, many of the questions in the pre- and post-questionnaires are taken from journal reading questions tested and asked many times in prior research by Tenopir & King. The application of multiple methods allows us to acquire and interpret data about the research questions from complementary perspectives. Individually, each method presents particular advantages and disadvantages in its ability to cover the richness and complexity of the use and utility of journal article component indexing. Thoughtful combinations of methods allows us to address limitations of one method with data gathered using a complementary method. This allows researchers to triangulate, or draw inferences or conclusions about the phenomena being studied, with more confidence than relying on a single method (Bryman, 2003; *Triangulation. World of Sociology*, 2001).

Both the reliability and validity are hampered, however, by the fact that participants interacted only with a prototype system. The size of the database was limited and some of the tables were illegible. The results must be interpreted with this in mind. Some participants answered questions based on what they believed the prototype system could become, not completely on their experiences with the instantiation of the system they used. Others were distracted by the limitations of the prototype and answered questions based on those limitations, which are not likely to be a part of an operational system. Additionally, the relatively small sample and its nature as a sample of convenience (discussed below) limit the ability to generalize from the data reported here. The results are, however, important in revealing the extent of use of tables and figures drawn from scholarly journal articles.

CSA, the sponsor of this project³, contacted institutions that would likely provide access to participants for this study. CSA's contacts at each institution then recruited individual researchers at their institution. The institutions selected were a mix of universities and research institutes located in the United States and Europe (Table 1).

	Universities	Research Institutes	Totals
United States	5	1	6
Europe	2	1	3
Totals	7	2	9

Table 1: Types and locations of sites used to gather data to investigate the impact of direct search of tables and figures drawn from scholarly journal articles.

³ The involvement of CSA in the project was limited to identifying institutions likely to participate, putting the researchers in contact with a librarian at each institution, and commenting on first drafts of the data collection instruments. The librarians at each institution recruited the participants. The researchers were solely responsible for all data collection and analysis. CSA agreed at the inception of the project to allow the researchers to disseminate results without any restrictions.

Each institution recruited between four and twelve scientists and researchers, for a total of sixty participants. This sample of convenience yielded a group of participants representing a cross section of science subject disciplines, geographic spread, and academic level. Although it is not random, the number of institutions and participants at each institution represents an adequate sample from which to draw meaningful conclusions.

Participants self-identified by providing their academic rank and/or job title to provide a rough indication of each participant's research experience. A plurality of participants identified themselves as either professors or researchers, denoting a high level of experience. The participant pool also included a large number of post-doctoral researchers. The next largest group was "students," who all held at least a bachelor's degree. Librarians, while not the primary focus of this project, represent an important constituency because they act as intermediaries and are often responsible for the acquisition, promotion, and training of users for all kinds of searching and abstracting and indexing systems at their institutions (see Table 2).

Rank/Title	Count	Percentage
Professor	17	28
Researcher	7	12
Post-doc	18	30
Student	11	18
Librarian	5	8
Unspecified	2	3

Table 2: Rank / Title of participants involved in investigation of the impact of direct search of tables and figures drawn from scholarly journal articles.

Participants also provided information on their highest-level academic degree. Two-thirds hold a PhD or MD. Masters degrees are held by nine participants and Bachelors degrees by nine participants. Two participants did not respond to this question (Table 3).

Degree	Count	Percentage
PhD	39	65
MD	1	2
Masters	9	15
Bachelors	9	15
Unspecified	2	3

Table 3: Highest academic degrees held by participants involved in investigation of the impact of direct search of tables and figures drawn from scholarly journal articles.

We asked participants to provide the disciplinary label with which they most closely identify. Various sub-fields of biology were grouped together as "Biology" and comprise a near-majority.⁴ The discipline occurring with the second-highest frequency is "Ecology". All others occurred at a frequency of five or fewer (Table 4).

⁴ We used a combination of information from two questions regarding participants' department/unit and their principal discipline. Responses combined into the biology category included Biology, Molecular Structural Biology,

Discipline	Count	Percentage
Biology	28	47
Ecology	8	13
Biochemistry	5	8
Library	5	8
Geology	4	7
Medicine	3	5
Oceanography	2	3
Forestry	2	3
Engineering	1	2
Entomology	1	2
Chemistry	1	2

Table 4: Principal discipline with which participants most closely identify. Participants were involved in investigation of the impact of direct search of tables and figures drawn from scholarly journal articles.

In summary, most participants hold a PhD or similar terminal degree; most are working in the life sciences; all are serious consumers of research and users of electronic search systems; and most are also producers of primary scientific research and journal articles. Research consumes the majority of the time of three-quarters of the participants, and over half of the participants spend up to half of their time on teaching. While we might have hoped for a somewhat higher proportion of senior researchers, the variety of the participants provides a wide range of perspectives with regard to expectations, needs, and uses for scientific information systems.

Findings

This section presents selected findings that emerged from this assessment of the impact on scientific work of providing a new, finer-grained means of accessing scientific journal literature. The findings presented here are based primarily on the data collect in the open-ended questions in the structured diaries and on the questionnaires as well as the data collected during the observations of component prototype use.

Finding relevant components and documents

Data collected during this project affirm that scientists overwhelmingly use digital versions of journal articles to support their research and teaching.⁵ They also usually perform most of their own searches when seeking journal articles and journal article components. Scientists are motivated to use digital systems because of the conveniences they afford, and wish to avoid the delays caused by traveling to the library or waiting for document delivery services, and wish to avoid obtaining poor quality representations of components, particularly photographs and other images, caused by photocopying or faxing. Seventy-five percent of the

Molecular Biology, Marine Ecology, Microbiology, Marine Science/Biology, Biochemistry, Cellular and Molecular Biology/Genetics, Ecology & Evolutionary Biology, Biological Science/Neuroscience, and Cell/Cellular Biology.

⁵ Note that this study focuses on searching to support teaching and research rather than activities associated with using journals to maintain current awareness.

participants reported that seventy percent or more of the journal articles they read⁶ were obtained or viewed in an electronic format; seventeen of sixty participants reported using electronic formats exclusively. Forty-two of the participants reported performing all of their own searches with seven of the participants performing 90-99% of their searches themselves.

Participants' current experiences revealed several issues illustrating current challenges to finding (and using) relevant components drawn from journal articles. They frequently referred to the important, direct, and problematic relationships between query term selection, results set size, and system features (such as search limiting and results filtering). While these are not recently discovered problems (e.g., Olsen, 1994; Lancaster, 1995; Institute for the Future, 2002c; Research Information Network, 2006), it is important to note them in order to ground the later discussion of the findings, which are presented in the context of scientists' experiences with contemporary search systems. Some noted general difficulties using Boolean search systems effectively, but others commented on the problem of selecting query terms at the appropriate level of specificity. It "can sometimes be very [difficult] to find the middle ground between too general a search – returning too many articles – and too specific a search, returning too few" (Post-doc, Ecology).

Results set size is another critical problem because of the cognitive load and stress it puts on the scientist: "the number of papers in a particular field are overwhelming" (Researcher, Biochemistry). Participants also recognize the difficulties caused by high-recall, low-precision searches: "Searches may yield large [numbers] of results – but not many relevant results" (Librarian). Finally, because of the challenge of finding a few relevant documents out of a large results set, scientists fear "finding out that I've missed a critical article that I should have found" (Post-doc, Ecology). Fear of missing significant items was also cited as a "main concern" of researchers due either to "inadequacies in a particular discovery service or in their own expertise" (Research Information Network, 2006).

Scientists recognize that search systems could be improved by providing additional tools to help them cope with large results sets, which are caused both by the query term selection problem and the ever-expanding universe of documents available in each database and scientific domain. When it's necessary to use a very general term such as "light" because there are no widely-used, more-specific terms, scientists need other tools to limit or filter the results in order to obtain reasonably sized results sets. "I'd like to see ways to filter results be more user friendly and prominent" (Student, Biology) and "I often want to exclude biomedical articles because of their relatively large [number]" (Post-doc, Ecology).

The scientists participating in this study provided confirmation of our expectations that there remains a consistent, unmet need for systems that yield higher precision searches coupled with seamless, universal access – ideally at no direct cost to the researcher – to high-quality artifacts, including journal article components like figures, tables, graphs, maps, and photographs. The scientists also understand that several factors interact in current systems, including increasing numbers of publications and articles, lack of correspondingly effective

⁶ In this study, the term *reading* is defined as "going beyond the table of contents, title, and abstract to the body of the article."

search and filtering tools, and the problem of selecting query terms that match the most relevant components and documents.

Eleven of the sixty participants provided information about using currently available systems to search for specific kinds of article components or objects, such as photographs, maps, figures, graphs, and tables. Specific systems used included Google Image Search, Google Maps, Google, Yahoo, AltaVista, MapQuest, VitalBook, PubMed, and PsychINFO. Three other kinds of components or objects participants searched for were *tests/measurement instruments*; *posters*; and *images*. Participants noted that using these systems to locate components is “difficult” (Professor, Medicine) and that the systems they used relied on “keyword searching [which is] not always the best method” (Librarian). Some also noted dissatisfaction with using general search systems, like Google, for finding images and other components because “results for academic articles are not good” (Researcher, Medicine) and that “in general, academic figures, tables, and graphs are not available to search using these services” (Professor, Engineering). Others had positive experiences (“Google Image is great for photos” - Post Doc, Biology), even if success was not immediate: “It took me a while to get used [to] such searches but then it became easier and I could find the relevant literature of my interest” (Researcher, Biochemistry).

Existing systems that support component searching are few in number, have significant limitations, and are infrequently utilized. These conditions present opportunities for further research and innovation by librarians and commercial service providers. There are three aspects at play here: first, general systems, like Google Image Search, are tuned neither in terms of their content (that is, what gets indexed, and with what level of quality) nor in terms of the tools available for conducting precise searching. They are not appropriate tools for scientists, even if they do sometimes provide useful results. Second, the tools that exist are not constructed to discriminate between a wider variety of components, such as graphs, tables, maps, and photographs, at the level of detail and quality that is often required by a scientist: these systems cannot distinguish popular images from those more typically of scholarly interest: “Google maps were very helpful for me in the past, but they are not part of the scientific articles” (Researcher, Biochemistry). Third, there is still a general lack of awareness of the availability and suitability of tools for searching at finer levels of granularity, like the article component level. “I have no idea how to search fig[ure] or table, etc.” (Post Doc, Biology) was one comment a participant made prior to using the component prototype. “I am not aware of any online services that allow search for Figures, Tables, Graphs..., but I would appreciate this service” (Researcher, Biochemistry). “My object searches to date have been limited, primarily because I'm not aware of search engines that effectively locate figures, graphs, or tables” (Post Doc, Ecology).

The data suggests that disciplinary differences affect how specific participants approached the prototype and the specific features they found most useful. Certain disciplines rely more on certain kinds of components than others: geologists (and some biologists and ecologists), for example, rely on maps and want to search by location: “Geology is based on maps.... there is no central place that indexes them. ...the ability to search by location would be helpfu” (Researcher, Geology). “Finding articles with analytical data about geological samples (geochronological information or geochemical information) that also include maps

where the samples were collected would be valuable” (Professor, Geology). Some researchers rely upon high-quality images and photographs: “For research, clarity of data [in images] is important to understand the study I am searching for” (Post Doc, Biology). However, searching by latitude and longitude values was not cited as a highly desired feature (see also Borgman et al, 2005).

Other participants were intrigued by the possibilities of having more direct access to the data expressed in tables and graphs to support data compilation and comparison. Some participants noted that some kinds of information are typically represented in particular types of components, while other researchers spoke more generally about the utility of various representations: “Directly relevant data in a summarized form (table, figure, etc) are very difficult to find quickly” (Researcher, Ecology). “I often look for information on specific bacterial strains or zoological specimen number; this information is often tabulated, even if not mentioned in the article text. Sometime these number[s] appear on graphs as well - less often on maps” (Post Doc, Biology). “Finding graphs, or tables, featuring particular measurements/metrics (e.g., height, biomass), and/or particular species (e.g., Red Oak). These details are often not included in title, abstract, keywords, etc.” (Post Doc, Ecology). “Finding data (tabular or graphical) that met some criteria in order to compare with models. Generating new data is expensive & time consuming. Finding existing relevant data is therefore valuable” (Post Doc, Biology).

Assessing relevance and interpreting components

Scientists anticipated gaining faster and more precise understanding of the work reported in other papers by direct examination of the components embedded in other articles, like images, figures and tables, when triaging, or determining the relevance of, individual articles (Bishop, 1998; Bishop, 1999; Bishop et al, 2000; Institute for the Future, 2002b). Direct access to these components “would allow me to faster understand/discern the statistical analyses employed instead of having to read the [materials and methods] section (I do a lot of multivariate statistics)” (Post Doc, Ecology). Another said “I work on developing a global understanding of carbon exchange between ocean and atmosphere. Being able to find visualisations of other attempts at regional or global synthesis would save me a lot of time” (Professor, Oceanography). Another stated that seeing “the thumbnails [is] often sufficient, that once I see those I know whether or not the item is worth pursuing” (Post Doc, Geology). The enhanced abstract, which includes thumbnails of all of the tables and figures drawn from the article, allows for fast characterization of entire articles by making it possible to “distinguish ... primarily theory papers and those that present data” (Post Doc, Biology).

The quality of the component returned by the prototype is critical to support accurate relevance assessment, particularly for researchers depending upon photographs provided in high resolution in printed journals. “Some journals are not available online and have to be ordered through the librarian and the re-prints or photocopies are horrible for any kind of analysis” (Post-doc, Biology). This is a problem noted by other researchers: “The ‘last mile’ of the process which actually delivers the document or other source that has been searched for is the [researchers’] focus of concern, with lack of access to journal articles because of a

subscription barrier being the most frequently-expressed difficulty” (Research Information Network, 2006).

However, scientists cautioned against using components drawn from articles without considering the context provided by the components’ embedding articles. When asked in the post-search questionnaire whether these components could be used on their own or if the article was also necessary to understand the component, only 9 (20%) of the 46 valid responses agreed that the objects were useful on their own, and 37 (80%) indicated that access to the objects’ corresponding article is required for the object to be useful. Several participants alluded to the dangers of relying upon de-contextualized tables and figures. “Using just the annotation of the figures can be misleading in some cases (e.g. a generic protein name in the annotation may mean the isoform from human or from yeast)” (Professor, Biochemistry). A researcher in medicine wrote:

“Kind of hard to imagine as a table, figure, or graph is useless to me without its context. I would need the rest of the paper to figure out what the objects really means and more importantly, should I trust what it says. So I guess I would need to already be familiar with the article. An object search might be valuable in this situation.”

One participant stated more forcefully: “In general, tables and figures do NOT stand alone. If you are unfamiliar with a paper, having the tables and figures by themselves does not give you sufficient information to interpret them” (Post Doc, Biology).

The clearest expression of the risk of de-contextualized tables and figures came from a participant who presented a specific example that also suggests that changes to scientific writing are necessary to accommodate this innovation:

“I feel that it's downright dangerous to present tables outside of the context of the article, at least until researchers realize that their data may be handled in this manner and adjust the content of tables and figures appropriately (e.g. including "but see figure 3" in the caption when presenting two figures of contradictory findings, so that one figure cannot be considered in the absence of the other data).” (Post Doc, Ecology)

The quality of the captions associated with the tables and figures was noted by several participants. One wrote that “sometimes they [tables and figures] can be useful on their own, but in 95% of the cases the full text is needed, since normally the information contained in the captions does not allow a complete understanding of the illustration” (Student, Biology). At the time the participants were testing the system, many table and figure captions were incomplete (participants frequently noted this problem with the prototype), and it is possible that the preceding comment referred to the unintended truncation of captions in the prototype, but it is also possible that the participant is indicating that even “complete” captions provide insufficient information to completely understand the tables and figures in the vast majority of cases.

Several commented that tables and figures can be useful on their own, but all of these participants noted explicitly that the full-text of the article needs to be available as well for consultation: one participant called links to the articles “essential.” (Post Doc, Oceanography) Participants noted in particular that they might need the full article to provide sufficient context for understanding the tables and figures found: “In most cases, the supporting article should be used as well to put the research in context.” (Professor, Engineering) One participant noted that in cases where the quality of the reproduced tables and figures is poor, the full article is required to actually see the tables and figures: “in particular Greek letters and mathematical symbols don’t translate well, and often the tables are not readable from the version in [the prototype] and I refer to the original article.” (Professor, Geology) One participant noted that “the real answer to this question is somewhere between the two extremes. Some tabular data are self-explanatory, but I suspect that it is more common to need to read about the data before understanding its significance.” (Researcher, Geology)

A pair of participants wrote about specific sections of journal articles that would be most useful in providing context for information presented in stand-alone figures and tables, which suggests another intermediate level of disaggregation of the journal article, which is similar to Kircz’s (2002) conceptualization. “For a complete understanding/interpretation of the data, AT LEAST the math/met[hod] part of the article is absolutely essential.” (Student, Chemistry) Another wrote “As long as some information is provided, perhaps not the entire article may be necessary. The question becomes what is the minimum info necessary to bring a figure and/or table into context with the text.” (Professor, Biology)

Three participants commented that particular types of objects may be better able stand on their own, apart from the embedding context of the articles from which they are drawn. “Maps or photographs could be useful without the complete articles. (Post Doc, Ecology) Another noted a special case of using objects taken from articles written in a foreign language: “In some cases our researche[r]s have found articles useful for the tables even when the text is in a language they can't read.” (Librarian)

Participants also alluded to potential changes in writing practice: first, scientists may have to consider how they write their captions in order to provide sufficient context so they can better stand alone. Second, it may be possible to provide sufficient context with something less than the entire article; a section of the article, such as the methods, instead. Finally, some participants noted that certain types of objects, particularly maps or photographs, might be better able to stand alone outside of the context of their embedding article.

Component reaggregation and use

Participants identified a range of potential uses for the items found by using a system supporting direct search for and retrieval of article components. First, the components are used directly, within the context of the prototype, to make relevance judgments about documents or components. After relevance is established, information gleaned from disaggregated components may be incorporated into new, fixed text documents, into

presentations of various kinds, into other contexts and assemblages, or used as a basis for a variety of comparisons.

An obvious use of disaggregated components is in support of creating new, fixed documents related to research, such as original journal or conference papers, research proposals, meta analyses, technical reports, and review papers. This kind of use was analyzed extensively by Bishop (1999). One participant cited several possibilities: “In research (meta-analyses, review of relevant data on specific topics, and guidance in table/figure design)” (Post Doc, Ecology).

In other cases, components are incorporated into text or multimedia documents used to support performative activities, such as presentations in the classroom, at conferences, or at job interviews. A professor of biochemistry uses components found by using traditional indexing systems to “incorporate current publications into classroom lecture.” One motivation for incorporating components from articles into presentations is that “in teaching, visual representation is very important.” (Post Doc, Biology)

Finally, participants sometimes mentioned how journal article components could support the development of new objects or assemblages including software, new formulae, models, simulations, data compilations and hypotheses. “Having explicit presentations of data, such as in table form, allows me to quickly use that paper for downstream applications such as formula generation or system modeling” (Student, Biology). Another participant noted that “It would be very useful when trying to make compilations of existing data (especially tables)” (Post Doc, Geology). One professor of geology mentioned using “figures for modification and addition in research.” A post-doc in biology said “Generating new data is expensive & time consuming. Finding existing relevant data is therefore valuable.”

Participants also identified a number of other ways in which disaggregated article components can be used other than being re-fixed in documents, presentations, or other objects. Some of these include using tables and figures for making a variety of comparisons, a use not before noted by Bishop or other researchers. Many of these comparisons had to do with checking one’s own work results, methods, or instruments against the work, results, methods, or instruments of other scientists or placing one’s own work into the context of research in the discipline. “Figures: when comparing experimental setups to my own” (Post Doc, Ecology). “Finding data (tabular or graphical) that met some criteria in order to compare with models” (Post Doc, Biology). “Search to see if similar controls were used and also ... results seen by other groups.... confirm/assure if your experimental approach is correct [compared] to that followed by others” (Post Doc, Biochemistry). “... know if someone has generated data that might be similar or complementary to my own” (Professor, Biology).

Implications

Implications from this research apply to systems design, scholarly communication, scientific practice, information providers and professionals, and future research. The component prototype evaluated here is but one step toward future sets of sociotechnical arrangements in

support of scientific communication, and not a stable endpoint. “Increasingly, the boundary between resources themselves and discovery services is a permeable one, and this trend seems likely to continue as new forms of content aggregation are developed” (Research Information Network, 2006). The journal article component index prototype is by design a document disaggregation system. It adds value to the disaggregation process by adding descriptors to each disaggregated component, through the effectiveness of the indexes built upon the extracted data, and the quality of the interfaces available to support human use of the index. The implications below are based upon the current project as well as upon previous research (e.g., Bishop et al, 2000). In contrast to Bishop’s (1999) assertion that “it would be highly useful to search or view individual article components, but only in particular situations which, moreover, may not arise very frequently” (p.272), it appears that situations in which searching for tables and figures is useful are more common than previously understood.

Implications for systems design. A number of participants expressed frustration with the way the system presented results as a list of distinct journal article *components*. This was most troubling to users when several adjacent components in a result set were all from the same journal *article*: the interface provided no obvious grouping of these related items. One participant said: “I found it annoying that each article returned SO MANY tables and figures – it slowed my progress through the search results, and would really be problematic if the database was extended across many years. A ‘collapse by article’ option would be nice...” (Post Doc, Ecology). In addition to being a request for a specific system capability, we interpret this as an effect of the novelty of a system that searches for and presents a list of journal article *components* instead of a list of journal *articles*. Our sense is that users sometimes conflated searching for components with searching for articles. Indeed, the interface supports this conflation because of the ease with which a user can move from a detailed display of a single component and a display of the representation of the article (the enhanced abstract) containing a set of thumbnail images – representing individual components – extracted from the article. Users’ level of confusion was not so high as to make it impossible for them to proceed with using the prototype system.

New system features may have a typical trajectory: when introduced, they are exotic and perhaps poorly understood. They may not even be noticed, much less used, because their existence is not yet expected. After a period of time, the features may become prosaic as more systems incorporate it (Bishop, 1999; Institute for the Future, 2002b). An innovative system such as the journal article component indexing system needs a powerful, flexible, and transparent interface to be effective for scientists and researchers. This is particularly important when a vast majority (82%) of the participants indicated that they performed 100% of their own searching, and only two participants reported searches performed on their behalf by librarians. Two important implications to be drawn are that, first, *interfaces must be transparent* to allow people who are domain experts but not information retrieval experts to understand how they work and provide cues about modifying search strategies when the results provided from searches are do not meet researcher expectations. “It is critical that new systems and novel system features stand out;” outreach and training efforts are necessary because it’s “important to point out those features that are likely to be novel or unique and to explain them in terms of features users are likely to be familiar with already” (Bishop, 1999)

(p.273). This point is also noted by the Stanford e-Journals project (Institute for the Future, 2002b). Second, most searching by domain experts occurs in locations like offices and laboratories, far from where information retrieval experts, such as librarians, work. Thus, opportunities for direct communication between the domain and information retrieval experts are limited, and *domain experts need information – via interface cues and conventions where possible and through integrated, contextualized, and “smart” help systems – to facilitate successful information search and discovery.*

Other participants suggested capabilities that are beyond the scope of the current prototype, but are feasible, and could be added to the prototype in the future. There was some interest in making the contents of journal article tables searchable and to support the extraction of data from tables found by using the journal article component index. Two other specific examples were “To search for articles/figures in x format (for example: genetic distance) would yield many relevant results in related fields...” (Researcher, Ecology) and “I would like to pick a subject, such as "how many insects are located in a geographic region. It would be nice to be able to have a pie chart show up to give me that data!” (Researcher, Biology) Similar capabilities have been made part of other system implementations (e.g., Phelps & Wilensky, 2001).

Implications for scholarly communication. This study raises the issue of *granularity* in the context of the efforts of information scientists to control and improve access to the expanding universe of digitized information. Consider the varying levels of granularity implemented by different kinds of indexing systems: conventional journal indexing systems are designed to locate and return individual articles; online public access catalogs in libraries are designed primarily to locate books, journals and journal issues (but not individual articles within journals); full text searching systems operate at a word or phrase (adjacent word) level of granularity. This study examines information access at the level of particular journal article components (figures and tables).

Participants also noted an important benefit provided by high-quality component search systems: faster and more efficient searching, with smaller, more precise results sets (addressing a major disadvantage of the kinds of search systems currently available as noted earlier). Figures, tables, and other components are considered to be condensed summaries of an article’s content, making it possible for scientists to “find documents according to the data analysed in the experiments of an article instead of having to go through countless [number] of abstracts. (Post Doc, Biochemistry), thus allowing the scientist to “save time by avoiding acquiring full articles and reading more than needed” (Professor, Biochemistry). Librarians noted the potential of using component searches to directly answer reference questions. One looked forward to being able to “find data not listed in standard handbooks” or to locate “seismic maps.” (Librarian) One participant expressed skepticism that searching for some kinds of data would even be possible: “Since there is no set norm to represent data points, I am not sure how I could/would search for Tables, Graphs, Maps, etc.” (Professor, Biology)

Participants recognized the potential of the component prototype to have a positive impact on teaching and in creating other kinds of presentations. Direct incorporation of components found into presentation software, such as PowerPoint for lectures, talks, and presentations –

both in the classroom and to other audiences, such as at conferences or job interviews – was frequently identified as a situation in which searching for specific article components like tables, figures, graphs, etc., would be valuable. Motivations include showing students the “main point of a paper” or “what microorganisms look like” (Post Doc, Biology). When researchers have a specific point to illustrate in a lecture or presentation, they want to be able to find a relevant figure, map, photograph or table without (1) reading abstracts, (2) obtaining the most promising articles, and (3) examining and extracting the relevant components. In some cases, particularly related to teaching, they are looking for particular figures and tables they have seen before: they often recall particular objects or their characteristics, but not the title, author or source of the corresponding article. Direct searching for figures and tables has promise to make this process more efficient.

Implications for scientific practice. First, participants were optimistic that this kind of tool, if nothing else, would save time in conducting the search and discovery process in support of both teaching and research. One participant noted that having direct access to journal article components “would greatly improve the searching process by limiting the results to the most significant matches” (Professor, Entomology). Another significant impact was that users were sometimes able to find data that they would not have found using conventional indexing systems. A post-doc in ecology wrote that the journal article component index “changed the type of information that I searched for – I focused on specific DATA rather than general ideas or theories.” Finally, some participants felt that this kind of system supports learning, by researchers, of new skills and methods, including how to effectively present scientific results in tables, figures, graphs, and other components. One participant noted that direct access to tables and figures can be a source of models for improving one’s own presentation of results through the transfer of knowledge about conventions and innovations in information visualization. For one participant, the prototype “...made me think about different ways that data is conveyed...I’ll design my own future graphs and figures to better ‘stand alone’ as a result.” Another saw direct access to journal article components as a means for keeping aware or abreast of “newer/emerging science for educational purposes” (Researcher, Biochemistry).

Increasing use of discrete, de-contextualized journal article components also has implications for scientific writing and dissemination of research. As some participants noted, and has been argued previously (e.g., Kircz, 2002), discrete components need to be constructed by the researcher as if they might stand alone, outside of the traditional framework of a scholarly journal article. It is possible that, over time, the process of designing and writing scholarly articles would evolve to support increased disaggregation of journal article components.

Implications for information providers and professionals. Librarians and other information professionals, including information systems providers, should develop improved techniques for reaching out to communities of users in order to increase the knowledge and skills of the end users, and promote new, innovative capabilities to the user community, who are most likely searching alone. This is most important in situations where the target end users are encountering a system with features and capabilities that they are unlikely to have seen before or are unlikely to anticipate are available. Information professionals and information providers should also continue to develop component search and retrieval systems targeted at

the formal and gray scientific literatures. Improved human and automated techniques are needed to provide high-quality indexing of components, and tools (e.g., search and retrieval interfaces) designed to support search for and use of journal article components. In addition, this study suggests that disciplinary differences and, potentially, differences in research approaches within a single discipline can affect users' reaction to and use of innovative information discovery and retrieval systems. Information professionals and providers should consider, for example, how previous research into discovery and use of geospatial information can be applied to the design of a system such as the component prototype (Hill et al, 2000; Borgman et al, 2005). Other previous research (e.g., Palmer, 1991) has demonstrated that personal and contextual differences (e.g., cognitive style, task role) can affect information system use.

The data collected indicate that the participants are pragmatic about intellectual property issues. As long as the other people using the components they have published give appropriate attribution, they are satisfied. However, since journal publishers were not participants in this study, their points of view on the intellectual property issues are not represented.

Implications for future research. Finally, a number of interesting questions have emerged from this study that can act as an agenda for future research. First, we would like to engage directly with scientists from these and other disciplines regarding their use of components in journal articles and other sources in order to better understand how they are pulled apart and reaggregated, what tools are used to support those activities, and what kinds of tools are needed to facilitate use and re-use. What we gleaned in this project about this topic was limited to some extent by the project's focus on the evaluation of a specific system prototype. Performing a series of interviews and observations unconstrained by the presence of any specific system would likely reveal other kinds of components, perhaps completely separate from formal or gray literature, that scientists routinely dis- and re-aggregate. We would expect to find indications of this, particularly in light of the rapid and continuing evolution of "e-research," distributed collaboratories, cyberinfrastructure, grid computing, wikis, a wider range of open access literature, ontology development, and institutional and other domain-oriented repositories. Second, conduct within-discipline case studies in support of understanding the similarities, differences, and conventions that play within and between domains. As noted above, the data suggest that there may be important differences in the patterns of information and systems use both between different disciplines but also within disciplines, depending upon the individual researcher's specialty. Finally, we would like to identify other emerging or established systems allowing user interaction with components. For example, perform contextual studies of the use and non-use of "multi-valent" information systems, such as handbooks containing "live tables" and "live equations" to assess the utility of those systems.

Conclusions

This paper reports findings from a study of scientists' use of a system that directly indexes journal article components. The paper assesses the impact of this innovation on scientists' use of the literature as a part of scientific practice. Quantitative and qualitative data were

collected from sixty participants, most of who are active researchers, from nine organizations in Europe and the United States. Data was gathered using pre- and post-search questionnaires, observations of searching, and structured diaries. Although the participants constitute a self-selected sample, and that means they are more likely to be active online searchers, most are not experts in the use of CSA's other systems. We therefore assume that these participants are typical researchers and scientists in universities and research institutes. They use a variety of other proprietary and public databases in their work, including Web search engines, and often prefer using other systems such as Web of Science or PubMed. All of the participants are active consumers of journal articles and most, as active researchers, are also producers of journal articles. Many participants also have some teaching or other responsibilities.

Lancaster (1995) suggests that "information may be no more accessible today than it was fifty or even a hundred years ago... because of the growth of the literature, its scatter, and the increasing twigging of knowledge" (p.13). Systems supporting indexing of journal article components, such as figures and tables, are regarded as promising tools for *finding relevant components and documents* by scientists, particularly in their potential to save time by providing higher-precision search results than are available using conventional indexing and abstracting systems. Few scientists are currently accustomed to searching directly for these kinds of components, and those that currently search for components are using tools such as MapQuest and Google Image Search, which are not designed to support searching for specific kinds of components solely within the scope of scholarly journal articles.

Scientists are encouraged by the potential for searching the scientific literature at a finer level of granularity – at the level of disaggregated journal article components – because tables and figures are often considered to be concise summaries of the main points of journal articles. Such components summarize data, relationships, methods, sites, organisms, and variables in a way that textual abstracts cannot. Scientists find that free text searching of abstracts or full-text is frustrating because results sets often include articles that mention one or more query terms somewhere in the article, even when those terms are not central to the article's purpose. The data gathered here also reveals, however, that users move quickly and unconsciously between modes of use, sometimes focusing on search and retrieval of the components and sometimes focusing on search and retrieval of the documents from which the components have been disaggregated. Although they did not express it in these terms, participants see that tables and figures indexing may be an aid to both precision (reducing the number of irrelevant items in a search set) and recall (allowing them to find relevant tables, figures, or complete articles that they would not be able to find otherwise).

Scientific journal article components such as tables and figures are seen by scientists as encapsulating the most salient information in the article: the data, relationships between variables, maps, and images of organisms and test apparatus. Direct access to these components aids *assessment of component and document relevance* because tables and figures are often one of the first parts of an article scanned or read by a researcher after obtaining the complete text of the article. In cases where a conventional textual abstract display is not sufficient to make a final relevance judgment, the scientist must take additional steps to acquire the document – electronically or in hard copy – and, at a minimum, scan the

paper in order to make a final determination of its relevance. The presence of individual figure and table components in the results set and providing a collection of thumbnails in the enhanced abstract display brings additional, highly salient information to the user prior to examination of the article full text. Many participants were skeptical, however, about complete reliance by readers upon decontextualized journal article components: it is too easy to misinterpret or misunderstand a table, graph, map, photograph or figure without access to additional contextual information contained in the full text. One participant called presentation of decontextualized components “dangerous” because scientists, as authors, create their tables and figures within the context of an embedding article: the components are not designed to be understandable as stand-alone objects.

The participants in this study responded positively to the presentation of journal article components in the results set (Figure 1), and the ability to *read, interpret, and evaluate journal article components* immediately. The full record associated with each indexed journal article component usually presents a higher resolution image, the full caption, and descriptors assigned to that specific component in addition to bibliographic information about its embedding article (Figure 2). They also responded positively to the display of the enhanced abstract (Figure 3), which included hyperlinked thumbnail images of all components associated with that article because even thumbnails help scientists distinguish data-oriented articles (those with many tables, graphs, maps and detailed figures) from those that are not data-oriented. The collection of thumbnails is presented along with the abstract, descriptors assigned to the article, as well as other standard bibliographic information. Access to the full text of the article is still necessary to support use of the prototype as an article-locating system or to access the article text that fully contextualizes the figure or table. Many participants voiced complaints about the processes currently used by libraries to link from abstracting and indexing systems to the full-text of licensed materials: these linking systems are considered cumbersome, confusing, involve multiple steps, and usually result in the spawning of additional Web browser windows. While participants noted several technical problems with the prototype, such as truncated captions, poorly-rendered tables, and too many images available only as thumbnails, such problems appear capable of being resolved in future system iterations.

The prototype journal article component index the participants evaluated performs, by design, *document disaggregation* on the user’s behalf. We did learn that scientists disaggregate journal articles acquired through more conventional systems as part of their standard practice. Scientists will use various tools including “screen prints” or advanced Portable Document Format (PDF) tools to pull components of interest out of journal articles that are available on the Internet, often in order to embed components in other assemblages such as those created to support presentations. The scientists we worked with overwhelmingly stated that they believed that this activity was protected by the fair use doctrine, assuming that the disaggregator properly attributed his or her sources. The scientists also were comfortable with others disaggregating their articles as long as their article was clearly attributed as the source of the component.

Disaggregated journal article components are *reaggregated and used* in many ways by scientists. In some cases, participant comments projected how they could imagine using

components in the future since our interactions with them were concurrent with or just after their introduction to the components index under evaluation. However, in other cases, participants identified types of reaggregation and use that are already part of the standard practices of scientists and researchers. Examples of use include retrieving and using images for teaching, research, and presentations: some wanted to easily illustrate concepts, organisms, and methods with graphics from the current research literature; some wanted to extract a high quality image to put into a presentation with standard presentation software; others would use the prototype index to find images, but would refer to the complete article before using the image. A frequently cited potential use was facilitate comparisons between their own work and the work of other scientists because a journal article component index “allows one to quickly identify relevant data for comparison and context to one’s own research:” tables and figures from other papers could be used to answer the questions such as “are my results consistent if I am using the same laboratory method as another researcher?” or “is my way of presenting results as effective as other researchers?” Another kind of use was to more effectively support meta-analysis such as that needed in review articles or in developing an overview of a topic.

Acknowledgements

This work was funded in part by CSA, developer of the prototype index used by the scientists who participated in the evaluation reported here. Participants in the project included Margaret Casado, University of Tennessee Libraries; Donald W. King; Bobbie Suttles, Center for Information Studies; and Alison Connor and Kelli Williams, Graduate Assistants. We thank Dania Bilal, Kimberly Black, Geof Bowker, Lorraine Normore, Boyd Rayward, and graduate students and faculty from the University of Tennessee’s College of Communication and Information for constructive comments on a draft of this paper. We also thank the sixty researchers who volunteered to participate in this study.

References

- Beheshti, J. (1989). A longitudinal study of the use of library books by undergraduate students. *Information Processing and Management*, 25(6), 737-744.
- Bishop, A.P. (1995). Scholarly journals on the net: a reader's assessment. *Library Trends*, 43(4), 544-570.
- Bishop, A.P. (1998). Digital libraries and knowledge disaggregation: The use of journal article components. *Proceedings of the third ACM conference on Digital libraries*. Pittsburgh, Pennsylvania, United States: ACM Press, 29-39.
- Bishop, A. P. (1999). Document structure and digital libraries: how researchers mobilize information in journal articles. *Information Processing and Management*, 35(3), 255-279.
- Bishop, A. P., Neumann, L. J., Star, S. L., Merkel, C., Ignacio, E., & Sandusky, R. J. (2000). Digital Libraries: Situating Use in Changing Information Infrastructure. *Journal of the American Society for Information Science*, 51(4), 394-413.
- Borghuis, M., Brinckman, H., Fischer, A., Hunter, K., van der Loo, E., ter Mors, R., Mostert, P., & Zijlstra, J. (1996). *TULIP: Final report*. New York: Elsevier Science. Retrieved 11 February 2007 from: <http://www.elsevier.com/wps/find/librariansinfo.librarians/tulipfr>.

-
- Borgman, C. L., Smart, L. J., Millwood, K. A., Finley, J. R., Champeny, L., Gilliland, A. J., et al. (2005). Comparing faculty information seeking in teaching and research: Implications for the design of digital libraries. *Journal of the American Society for Information Science and Technology*, 56(6), 636-657.
- Bryman, A. (2003). Triangulation. In M. S. Lewis-Beck, A. Bryman & F. T. Liao (Eds.), *The SAGE Encyclopedia of Social Science Research Methods*. Thousand Oaks, CA: Sage.
- Burbules, N. C., & Bruce, B. C. (1995). This is not a paper. *Educational Researcher*, 24(8), 12-18.
- Demner-Fushman, D., Few, B., Hauser, S. E., & Thoma, G. (2006). Automatically identifying health outcome information in MEDLINE records. *Journal of the American Medical Informatics Association*, 13(1), 52-60.
- De Waard, A. (2005). Science publishing and the semantic web, or: Why are you reading this on paper? Paper presented at the European Conference on the Semantic Web, 2005, Industry Forum, Heraklion, Greece. May 29 - June 1, 2005. Retrieved February 28, 2007, from <http://www.cs.uu.nl/people/anita/papers/deWaardECSW2005.pdf>
- Eason, K., Yu, L., & Harker, S. (2000). The use and usefulness of functions in electronic journals: The experience of the SuperJournal project. *Program*, 34(1), 1-28.
- Entlich, R., Garson, L., Lesk, M., Normore, L., Olsen, J., & Weibel, S. (1997). Making a digital library: The contents of the CORE project. *ACM Transactions on Information Systems*, 15(2), 103-123.
- Florance, V., & Marchionini, G. (1995). Information processing in the context of medical care, *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval*. Seattle, Washington, United States: ACM Press.
- Friedlander, A. (2002). *Dimensions and use of the scholarly information environment: Introduction to a data set*. Washington D.C.: Council on Library and Information Resources. Retrieved 11 February 2007 from: <http://www.clir.org/pubs/reports/pub110/contents.html>.
- Hart, R. L. (1998). The relationships between work roles and information gathering of the faculty at SUNY, college at fredonia. *Library & Information Science Research*, 20(2), 163-185.
- Hill, L. L., Carver, L., Larsgaard, M., Dolin, R., Smith, T. R., & Frew, J., et al. (2000). Alexandria digital library: User evaluation studies and system design. *Journal of the American Society for Information Science*, 51(3), 246-259.
- Hyldegård, J. (2006). Collaborative information behaviour—exploring Kuhlthau's information search process model in a group-based educational setting. *Information Processing & Management*, 42(1), 276-298.
- Institute for the Future. (2002a). *Reflections on branding and e-journals*. Prepared for the Stanford University Libraries' e-Journal User Study, January 2002. Retrieved 11 February 2007 from: http://ejust.stanford.edu/findings/interview_branding.pdf.
- Institute for the Future. (2002b). *Core scholarly information tasks and e-journal features: Expanded discussion*. Prepared for the Stanford University Libraries' e-Journal User Study, January 2002. Retrieved 11 February 2007 from: http://ejust.stanford.edu/findings/interview_coretasks.pdf.
- Institute for the Future. (2002c). *E-journal user study: Report of the second survey: The feature user survey*. Prepared for the Stanford University Libraries' e-Journal User
-

-
- Study, November 2002. Retrieved 11 February 2007 from:
http://ejust.stanford.edu/findings2/report_survey2.pdf.
- King, D.W., Tenopir, C., Montgomery, C.H., & Aerni, S.E. (2003). Patterns of journal use by faculty at three diverse universities. *D-Lib Magazine*, 9(10), n.p. Retrieved 11 February 2007 from <http://www.dlib.org/dlib/october03/king/10king.html>.
- Kircz, J. G. (1998). Modularity: the next form of scientific information presentation? *Journal of Documentation*, 54(2), 210-235.
- Kircz, J. G. (2002). New practices for electronic publishing 2: New forms of the scientific paper. *Learned Publishing*, 15(1), 27-32.
- Kling, R. & Callahan, E. (2003). Electronic journals, the Internet, and scholarly communication. *Annual Review of Information Science and Technology*, 37, 127-177.
- Kwok, K.L. (1990). Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems (TOIS)*, 8(4), 363-386.
- Lancaster, F.W. (1995). Needs, demands and motivations in the use of sources of information. *Journal of Information, Communication, and Library Science*, 1(3), 3-19.
- Liu, Y., Mitra, P., Giles, C. L., & Bai, K. (2006). Automatic extraction of table metadata from digital documents, *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries*. Chapel Hill, NC, USA: ACM Press, 339-340.
- Lu, X., Mitra, P., Wang, J. Z., & Giles, C. L. (2006). Automatic categorization of figures in scientific documents, *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries*. Chapel Hill, NC, USA: ACM Press, 129-138.
- Monk, A., Wright, P., Haber, J., & Davenport, L. (1993). *Improving your human-computer interface: A practical technique*. New York: Prentice-Hall.
- National Science Board. (2006). Science and Engineering Indicators. Two volumes. Arlington, VA: National Science Foundation (volume 1, NSB 06-01; volume 2, NSB 06-01A). Retrieved 11 February 2007, from <http://www.nsf.gov/statistics/seind06/>.
- Olsen, J. (1994). *Electronic journal literature: implications for scholars*. Westport, CT: Mecklermedia.
- Palmer, J. (1991). Scientists and information: II. personal factors in information behaviour. *Journal of Documentation*, 47(3), 254-275.
- Peek, R.P. & Pomerantz, J.P. (1998). Electronic scholarly journal publishing. *Annual Review of Information Science and Technology*, 33, 321-356.
- Phelps, T. A. & Wilensky, R. (2001). The multivalent browser: a platform for new ideas. Paper presented at the *Proceedings of the 2001 ACM symposium on document engineering*, Atlanta, Georgia, USA.
- Research Information Network. (2006). Researchers and discovery services: Behaviour, perceptions and needs. London: Research Information Network. Retrieved February 28, 2007, from <http://www.rin.ac.uk/files/Report%20-%20final.pdf>
- Schlieder, T., & Meuss, H. (2002). Querying and ranking XML documents. *Journal of the American Society for Information Science and Technology*, 53(6), 489-503.
- Simpson, A. (1988). Academic journal usage. *British Journal of Academic Librarianship*, 3(1), 25-36.
- Stewart, L. (1996). User acceptance of electronic journals: Interviews with chemists at Cornell University. *College & Research Libraries*, 57(4), 339-349.
-

- Tenopir, C. & King, D.W. (2000). *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers*. Washington, D.C.: Special Libraries Association.
- Tenopir, C. & King, D.W. (2004). *Communication Patterns of Engineers*. Piscataway, NJ: IEEE Press.
- Tenopir, C., King, D.W., Boyce, P., Grayson, M., Zhang, Y., & Ebuon, M. (2003). Patterns of journal use by scientists through three evolutionary phases. *D-Lib Magazine*, 9(5), n.p. Retrieved 11 February 2007 from <http://www.dlib.org/dlib/may03/king/05king.html>.
- Tenopir, C., Nicholas, D., & Baker, G. (2006). Scatter and Decay: E-journal Usage Patterns. *2006 XXVI Annual Charleston Conference Issues in Book and Serial Acquisition*, Charleston, South Carolina, November 8-11, 2006.
- Triangulation. *World of Sociology*, Gale (2001). Retrieved 23 May 2006, from xreferplus. <http://www.xreferplus.com/entry/4785970>
- Wilensky, R., & Phelps, T. A. (1996). *Toward active, extensible, networked documents: multivalent architecture and applications*. In *Proceedings of the 1st ACM international conference on digital libraries*, 100-108.