Figure and Table Retrieval from Scholarly Journal Articles: User Needs for Teaching and Research

Robert J. Sandusky

School of Information Sciences, University of Tennessee, 451 Communications Building, 1345 Circle Park Drive, Knoxville, TN 37996 Tel: +1 865-974-2785

sandusky@utk.edu

Carol Tenopir

School of Information Sciences, University of Tennessee, 451 Communications Building, 1345 Circle Park Drive, Knoxville, TN 37996

Margaret M. Casado

John C. Hodges Library, University of Tennessee, 135 John C Hodges Library, 1015 Volunteer Boulevard, Knoxville, TN 37996

This paper discusses user needs for a system that indexes tables and figures culled from scientific journal articles. These findings are taken from a comprehensive investigation into scientists' satisfaction with and use of a tables and figures retrieval prototype. Much previous research has examined the usability and features of digital libraries and other online retrieval systems that retrieve either full-text of journal articles, traditional article-level abstracts, or both. In contrast, this paper examines the needs of users directly searching for and accessing discrete journal article components - figures, tables, graphs, maps, and photographs - that have been individually indexed.

Introduction

Most previous research on scientists' use of electronic journal articles has focused on the use of entire articles for support of research activities including current awareness, writing papers and grant proposals, and investigating new research areas (Friedlander, 2002; Institute for the Future, 2002; Tenopir & King, 2004; National Science Board, 2006). This paper discusses scientists' needs related to searching for, retrieving, and using discrete journal article components for both their research and teaching activities, with a particular

focus on tables, figures, graphs, maps, and photographs. These findings are taken from a comprehensive investigation into scientists' satisfaction with and use of a prototype retrieval system that indexes individual tables and figures culled from scientific journal articles (Sandusky & Tenopir, forthcoming). This study found that researchers find figures and tables important to them both as independent objects (e.g., disaggregated from their context in journal articles and re-used as figures in presentations, lectures, and in support of other work; see Sandusky, Tenopir & Casado, 2007) and as a means to more efficiently identify relevant journal articles to support both research-related and teaching or presentation activities.

While some writers have argued that electronic documents provide opportunities for transforming publishing, indexing, and retrieval practices, there is little empirical research into scientists needs for reconfigured electronic documents or systems that support radically different publishing models. Kircz (2002) suggests that modularity should be the next dominant paradigm for scientific publishing. In Kircz' model, the traditional scholarly journal article is disaggregated into specific formal subcomponents that can be published and stand on their own, thus transforming scientific writing as well as publishing.

The direct interaction of users with journal article components has not often been the subject of empirical research studies. Notable exceptions include the work of Bishop and colleagues on the needs and behavior of the target users of a novel digital library supporting the retrieval of science and engineering journal articles via searching of specific journal article components, including figure and table captions and table text (Bishop 1999; Bishop et al, 2000) and Stelmaszewska & Blandford's (2004) study of physical library use. Based upon their research of how computer scientists read articles and assess relevance in physical libraries, Stelmasjewska and Blandford (2004) observed library users "flicking through journal pages, searching for images, figures, or formulae that could help the reader in evaluating the journal" and further noting that journal users "sometimes scan the paper's content, searching for pictures, figures, or formulae to make sure that it [the paper] fulfils their requirements." They noted these components as being of particular importance: the abstract, introduction, section headings, bulleted lists, summaries, definitions, pictures, and specific canonical sections such as methods, findings, future research, conclusions, and references. Remarking upon the consistency of their findings with those of Bishop (1999), they note that digital manifestations of journal papers typically lack convenient affordances supporting access to the journal article components that users rely upon to make relevance assessments. They suggest that digital library designers should seek to make such affordances available at the users' fingertips without proposing any specific design.

Voorbij and Ongering (2006) found that electronic journals and electronic versions of

articles do not inherently make assessment of the relevance of journal articles faster or easier. They note that about 20% researchers included in their survey agreed or strongly agreed with the statement "In order to judge the relevance of an electronic article, I need to print it out first." In their follow-up interviews with a subset of the researchers surveyed, some researchers felt they printed "too many articles and [threw] them away" (p.230), implying a need for improvements in the design of electronic journal and digital library systems to support online relevance assessment.

The few studies cited above provide evidence of the importance of journal article components (figures, tables, abstracts, methods sections, etc.) for supporting end user relevance assessments both in physical and electronic formats. This paper provides evidence that indexing and retrieval of figures, tables, graphs, photographs, and maps can make online relevance assessments easier in support of scientists' research and teaching activities. The journal article component prototype evaluated here gives end users direct access to disaggregated journal article components. The prototype provides an end-user search interface supporting construction of complex Boolean queries and selection of a variety of field limits and results filters. For example, the searcher can enter search terms and limit the string matching to figure or table caption text, or geographic, statistical, or taxonomic descriptors. The user can likewise limit the results returned to figures, tables, graphs, maps, and/or photographs. In addition, the prototype adds a collection of thumbnail images for each of an article's tables and figures to an otherwise traditional article-level abstract, further supporting online article-level relevance assessment (see Sandusky, Tenopir, & Casado, 2007 for illustrations of the prototype's interface).

Methods

Multiple methods were employed to collect both qualitative and quantitative data for this study. Data collection procedures and instruments were designed to provide insights into the potential impacts of the indexing and retrieval of journal article components on teaching and research. One of three researchers traveled to each 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 included distribution and collection of the pre-search questionnaire (to measure characteristics of the participants, prior knowledge, experiences with, and potential uses of journal article component indexing), an introduction to the journal article component indexing approach and the prototype, 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 presentation materials in order to minimize the differences in how the participants were introduced to the project.

At the end of the study, after participants turned in their structured diaries, a post-search questionnaire was distributed electronically. The post-search questionnaire contained some of the same questions as the pre-search questionnaire to examine how exposure to the journal article component prototype influenced participant's perceptions of the utility of journal article component indexing (for a more detailed description of the methods employed for this study, see Sandusky, Tenopir & Casado, 2007).

ProQuest CSA, the sponsor of this project, contacted institutions that would likely provide access to participants for this study. ProQuest 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 (5 universities and 1 research institute) and Europe (2 universities and 1 research institute). 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, if provisional, 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 research 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.

Results

This section presents the analysis of questions focused on identifying user needs for a journal article component indexing and retrieval system. First, participants' initial attitudes about the importance of being able to limit component searches by five specific components, for both teaching and research, are given. Second, data showing the importance of and participant satisfaction with the limiting categories provided by the prototype are provided. Other data are then presented that focus on more general user needs for supporting retrieval of full journal articles or components.

Prior to their use of the prototype, participants were asked to rank the importance of being able to search for specific components of journal articles. A pair of questions used Likert scales to elicit participants' rating of the importance of access to a high-quality search capability for various object types for supporting their teaching and their research. The six object types were tables, figures, graphs, maps, photographs, and other. Each question had its seven-point scale anchored by the terms "not important" and "absolutely essential." "Not applicable" was also an option available to the participants.

Table 1 provides a condensed summary of the participants' rankings of the importance of being able to search for particular journal article components in support of both their teaching and their research. Searching for tables, figures, graphs, and maps is considered more important for research; searching for photographs is about the same for teaching and research, and other types of objects are more important for teaching than they are for research (we did not gather information about what "other" types of objects the participants would find useful for teaching or research).

Table 1: Condensed comparison of responses for importance of six object types for support of teaching and research (N=60; from questionnaire administered prior to participants' use of the prototype).

Object Type	Importance: Teaching		Importance: Research		Difference in Mean
	Mean	Std. Dev.	Mean	Std. Dev.	
Tables	4.38	2.368	5.93	0.944	+1.55
Figures	5.07	2.324	5.87	1.033	+0.80
Graphs	4.93	2.313	5.68	1.195	+0.75
Maps	3.66	2.718	4.11	2.339	+0.45

Photographs	4.55	2.415	4.51	1.942	-0.04
Other	4.66	2.395	4.11	2.527	-0.55

It is difficult to see the differences between responses to the importance of particular journal article components shown in Table 1 without additional explication. In terms of each component type's importance for teaching, the pattern of responses is similar for figures, tables, graphs, and photographs: responses are clustered above the midpoint with far fewer responses at the opposite end of the scale. The response for maps was different: responses were spread across the seven-point scale and nearly one-quarter of the participants (13) responded "not applicable." For "other" types of components, only 7 participants provided a response on the seven-point scale; the remaining 53 participants marked "not applicable" or did not provide a response.

In terms of the importance of the importance of each type of component for research, the means of the ratings by all participants tended to be the same or higher (as shown in the last column of Table 1) except for the category "other." The responses also clustered more tightly as reflected in the lower standard deviations. The distribution of responses for figures, tables, and graphs were all quite similar, with the overwhelming majority of responses above the midpoint and no responses of either 1 or 2, and no responses of "not applicable." Photographs show a much greater spread in the responses, with 28% of the participants (17) providing responses below the midpoint. The responses for maps are essentially bi-modal, with 12 participants (20%) giving a rating of 1 (not at all important) and twenty-two participants (37%) giving a rating of 6 or 7. Nineteen participants gave ratings between 2 and 5 for maps, inclusive.

After the participants had used the prototype and completed and returned structured search diaries, an online summative questionnaire used seven-point scales to ask participants about both the importance of and their satisfaction with specific Tables & Figures index features (Table 2). Here, the questions did not distinguish between teaching and research. Participants rated feature importance on a scale where 1 was "not at all important" and 7 was "absolutely essential" and satisfaction on a scale where 1 was "totally dissatisfied" and 7 was "totally satisfied." They rated category limits in general, as well as these specific features for limiting searches: tables, figures, maps, graphs, photographs, predictive model, and free access. Forty-six valid responses were received.

Satisfaction ratings are quite similar for the "limits" features considered most important. For the limits categories given the lowest importance ratings (maps, photographs, predictive model, and free access), satisfaction with the implementation ranked higher than importance ranking.

Table 2: Condensed comparison of responses for importance and satisfaction with specific limiting features of the Tables & Figures Index (N=46; from questionnaire administered following participants' use of the prototype)

Feature	Importance		Satisfaction		Difference in Mean
	Mean	Std. Dev.	Mean	Std. Dev.	
Category limits in general	5.09	1.603	4.93	1.467	-0.16
Tables limiting	5.07	1.638	5.07	1.497	0.00
Figures limiting	5.26	1.555	5.09	1.518	-0.15
Graphs limiting	4.80	1.655	4.72	1.615	-0.08
Maps limiting	4.33	1.886	4.41	1.758	+0.08
Photographs limiting	4.26	1.937	4.46	1.669	+0.20
Predictive model limits	3.02	1.844	3.74	1.843	+0.72
Free access limits	3.78	2.107	4.59	1.600	+0.81

The pattern of response for the importance of category limits in general, tables, figures, and graphs are similar, with responses above the midpoint dominating. For maps, the most common response was the midpoint (10 participants), with 22 responses above the midpoint and 14 responses below the midpoint. For photographs, the midpoint was the most common response (10 participants), but 8 participants rated the importance of limiting by photographs at 7. For predictive model limits, the most common response was "not at all important," with 14 participants giving that response. Twelve participants rated predictive model limits above the midpoint, and 20 rated predictive model limits at or below the midpoint. Responses about the importance of free access limits were bimodal, with ratings of 1 (not at all important) and 6 each given by 10 participants; participants responses were fairly evenly distributed among the other five points on the scale.

The ratings for satisfaction with the limits implementation are quite uniform; this is likely the case because the implementation in the prototype was consistent. Satisfaction with predictive model limiting rated lowest among all of the options for setting limits, although 16 participants rated their satisfaction above the midpoint.

Turning to more general needs from indexing and retrieval systems, we asked participants to describe particularly favorable or unfavorable experiences with current indexing and retrieval systems during the past twelve months. Participants expressed a need for systems that

- yield higher precision searches
- use more powerful, flexible, transparent, and still usable interfaces

- provide seamless, universal access to consistently high-quality artifacts, such as journal articles, figures, and tables
- provide effective federated searching to support cross- and inter-disciplinary work
- provide a standardized solution for handling searches including diacriticals and symbols

The overwhelming majority (82%) of the participants also indicated that they performed 100% of their own searching; only 2 participants reported having searches performed on their behalf by librarians. While the nature of our sample may bias this number, an increase in end-user searching has important implications for systems and interface design, as discussed below.

Implications

Results pertaining to journal article component searching are combined here with findings about indexing and retrieval more generally because one of the most interesting patterns of behavior we observed was participants moving between pure component retrieval (e.g., seeking a figure to illustrate an organism for a lecture) and use of the prototype as a tool to conduct searches for complete journal articles (e.g., seeking articles about a specific protein). Some of the general issues participants identified with current article-level indexing and retrieval systems (noted above) are worth considering in the context user needs for component retrieval systems. In particular, scientists' needs for systems that (1) yield higher precision searches, (2) employ more powerful, flexible, usable, and transparent interfaces, and (3) provide seamless, universal access to consistently high-quality artifacts are of likely interest to systems and service providers.

HIGHER PRECISION SEARCHING.

When asked on the pre-search questionnaire about the importance of particular features of indexing and retrieval systems in general, an overwhelming majority of participants (53 of 60, or 88%) indicated that relevance of items retrieved was essential or absolutely essential to them (53 of 60 participants selected either 6 or 7 on a 7-point Likert scale). The level of satisfaction with the relevance of items retrieved from indexing and retrieval systems in general was markedly lower: 48 of 60 participants, or 80%, rated their satisfaction at either 4 (midpoint), 5, or 6. Although they did not express this need in terms of "precision" or "relevance assessment," it's clear that participants were interested in somehow being able to review a larger number of articles per unit of time than they were able to using standard online abstracting and indexing systems. The journal article

component prototype may provide more efficient searching of the literature by including additional, highly-valued journal article components in the article-level surrogate. A professor of biochemistry said “For research, abstract with figures and tables would allow much more efficient screening of publications and would thereby save time by avoiding acquiring full articles and reading more than needed.” A post doc in ecology talked about improved efficiency in an even more specific context: “Tables and figures ... would allow me to faster understand/discern the statistical analyses employed instead of having to read the M&M [methods and materials] section (I do a lot of multivariate statistics).” A post doc in biochemistry noted that tables and figures distill the data central to an article, thus conveying the article’s essence: “I would be able to find documents according to the data analysed in the experiments of an article instead of having to go through countless [numbers] of abstracts.”

USABLE, POWERFUL INTERFACES.

The degree to which end users perform their own searching (and do not rely upon librarians or others with indexing and retrieval expertise) has several implications for systems and interface design. First, interfaces must be transparent in order 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 do not meet researcher expectations. Second, most searching by domain experts probably 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, real-time, and/or “smart” help systems - to facilitate successful information search and discovery. Finally, librarians and other information retrieval experts, including providers, should develop additional techniques for reaching out to communities of users in order to increase the knowledge and skills of the end users, who are most likely searching alone. This final point may be particularly critical when innovative systems with features users are unlikely to anticipate, like the journal article component prototype, are introduced (see also Bishop, 1999).

HIGH-QUALITY, ACCESSIBLE ARTIFACTS.

Participants noted frustration when direct, free, and fast access was not available to full text of the articles they found. Lack of access to full-text was identified by different

participants as either an internal organizational problem ("The direct access to online articles via the state library does not function and much time is spent finding the paper and then retrieving it elsewhere e.g. JSTOR."), a part of the nature of the Web or particular search systems ("It would be very helpful in Web of Science citations would have a direct link to the paper."), related to their own work practices ("I usually want/need the material yesterday.... I have given up waiting for hardcover delivery of source material and work only electronically."), or a problem of the current publishing environment ("finding journal articles for which they request $40 for each article was very frustrating"). Access to the full-text artifact directly from the search system is problematic in current systems and implementation varies from institution to institution. Systems sometime link to the wrong item; the link is broken when it should work; the researcher's institution doesn't have the appropriate subscription; the remote server is down; or access may be via a multi-step process (e.g., SFX) that can confuse the user as multiple browser windows launch. The issue of access is not simply an issue of absence or presence: in some cases, while access to particular systems or collections is provided, the quality of the access is problematic. One participant noted that "the European Patent Office's database was slow" and another reported that "Georef can be very slow..."

One consistent theme heard from the participants in the current study was that the availability of high-resolution images of the tables and figures contained in a journal article made judging article relevance easier. The prototype had several issues related to artifact quality that frustrated participants including poor rendering of many tables (leaving them unreadable); truncated figure and table captions (undermining effective relevance assessment); and lack of larger, higher quality figure and table images for many returned components. The poor rendering of tables and truncated captions were bugs in the prototype that were scheduled to be corrected, but the lack of larger versions of images was systemic, due to either licensing restrictions from the publisher or lack of direct, digital feeds of tables and figures into the prototype database. This final issue undermines the usefulness of the system because the user must take several additional steps to obtain the complete article in order to assess component or embedding article relevance. Issues of access and component quality can also combine to frustrate users: "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."

COMPONENT CATEGORIES AND LIMITS.

The numbers reported in tables 1 and 2 show that the importance of the component categories varies, possibly because of differences in the importance of particular

component categories to different disciplines, or because of differences in individuals' approaches to their research. It is clear from these numbers, their distributions, and some of the open-ended responses we received that maps, and to a lesser extent, photographs, are different than figures, tables, and graphs. Geologists, some ecologists, and some biologists make frequent use of maps, but they are largely irrelevant to many other life scientists. Further analysis needs to be done to verify some of these suspected correlations between discipline and the relative importance of specific features. The results of continued investigation into the variations in responses within and between disciplines can be applied to the development of more effective systems and services, perhaps tailored on a per-discipline basis.

However, it seems prudent for service providers, in this case the developer of the journal article component prototype, to provide as wide a range of options for limiting searches as possible. Even though limiting by the components related to predictive models was rated the least important among the options supported by the prototype, 11 of the 46 participants responding rated it above the midpoint, and 6 of those 11 rated it as essential or absolutely essential.

Conclusion

The journal article component prototype evaluated in this study does not represent a paradigm shift in scholarly communication patterns and publishing, as Kircz (2002) suggests is possible. It is instead an important evolutionary step that allows great value to be added to conventional scientific articles at the indexing and retrieval stage of the scholarly publishing cycle. Journal article component retrieval, as implemented in the prototype, has potential to support previously unrecognized information practices and improve scientists' ability to make online relevance assessments at both the component and journal article levels.

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 Donald W. King; Bobbie Suttles, Center for Information Studies; and Alison Connor and Kelli Williams, Graduate Assistants.

References

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.

Friedlander, A. (2002). Dimensions and use of the scholarly information environment: Introduction to a data set. *Council on Library and Information Resources.* Washington D.C.: Retrieved February 11, 2007, from: http://www.clir.org/pubs/reports/pub110/contents.html

Institute for the Future. (2002). 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 February 11, 2007, from: http://ejust.stanford.edu/findings2/report_survey2.pdf

Kircz, J. G. (2002). New practices for electronic publishing 2: New forms of the scientific paper. *Learned Publishing,* 15(1), 27-32.

National Science Board. (2006). Science and Engineering Indicators. Two volumes. *National Science Foundation (volume 1, NSB 06-01; volume 2, NSB 06-01A).* Arlington, VA: Retrieved February 11, 2007, from: http://www.nsf.gov/statistics/seind06/

Sandusky, R.J., & Tenopir, C. (forthcoming). Finding and using journal article components: Impacts of disaggregation and reaggregation on scientific practice. Submitted for publication.

Sandusky, R.J., Tenopir, C., & Casado, M.M. (2007). Uses of figures and tables from scholarly journal articles in teaching and research. *Presented at the Annual Meeting of the American Society for Information Science and Technology,* Milwaukee, WI.

Stelmaszewska, H. & Blandford, A. (2004) From physical to digital: a case study of computer scientists' behaviour in physical libraries. *International Journal of Digital Libraries.* 4(2), 82-92.

Tenopir, C. & King, D.W. (2004). Communication Patterns of Engineers. *IEEE* NY: Wiley Interscience.

Voorbij, H., & Ongering, H. (2006). The use of electronic journals by Dutch researchers: A descriptive and exploratory study. *The Journal of Academic Librarianship,* 32(3), 223-237