White Paper on Electronic Journal Usage Statistics

by Judy Luther

Council on Library and Information Resources
Washington, D.C.

About the Author

Judy Luther is president of Informed Strategies, a consulting firm in Ardmore, Pennsylvania. Ms Luther specializes in researching and developing new products and companies, and introducing them to the library market. She has worked in the information field for 30 years and holds master's degrees in library science and in business administration. Her business experience includes 12 years in sales, sales management, and product development at Faxon and ISI. In her 13 years of library experience, she has held positions at Embry-Riddle Aeronautical University and Stetson University. Ms. Luther is active in program planning with the North American Serials Interest Group, the Society for Scholarly Publishers, and the American Library Association. Her articles frequently appear in *Information Today*, *Against the Grain,* and the *Charleston Advisor*.

ISBN 1-887334-79-3

Published by:

Council on Library and Information Resources
1755 Massachusetts Avenue, NW, Suite 500
Washington, DC 20036

Web site at http://www.clir.org

Additional copies are available for $15 per copy. Orders must be placed online through CLIR's Web site.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

First edition 2000
Second edition 2001

Contents

Acknowledgments

The author wishes to thank the Council on Library and Information Resources, and particularly Deanna Marcum, for supporting this project. This white paper required the involvement of both librarians and publishers who spent considerable time contributing their ideas, explaining developments and technology, and exploring the issues. They include Denise Davis at the National Commission on Libraries and Information Science; Martha Kyrillidou at the Association of Research Libraries; Tom Peters at the Committee for Institutional Cooperation; Tom Sanville at OhioLINK; Michele Newberry at the Florida Center for Library Automation; Rick Luce at Los Alamos National Labs; Arnold Hirshon at NELINET; Charles Lowry at the University of Maryland; Sue Phillips at the University of Texas at Austin; Jill Emery at the University of Texas at Arlington; Louise Green and Jim Mullins at Villanova University; Tom Gilson at the College of Charleston; Christian Boissonnas at Cornell University; Kent Mulliner at Ohio University; Fannie Cox and Weilung Liu at the University of Louisville; Stephen Wiberley and Debbie Blecic at the University of Illinois–Chicago; Judith Hiott at the Houston Public Library; Chrysanne Lowe and Taissa Kusma at Academic Press; Kristen Garlock and Kevin Guthrie at JSTOR; Jim McGinty at Cambridge Scientific Abstracts; Simon Inger and Kirsty Meddings at Catchword; Bernard Rous at the Association of Computing Machinery; Doug LaFrenier at the American Institute of Physics; Adam Chesler at Kluwer; Karen Hunter, John Carroll, and Marthyn Borghuis at Elsevier; Bridget Pairaudeau and Lloyd Fletcher at the Institute of Physics; and Dominic Martinez at MCB University Press.

Preface

Electronic journals represent a significant and growing part of the academic library's offerings. As demand for e-journals increases, librarians are faced with a new set of decisions related to acquisitions and services. Must libraries retain both print and electronic copies? Is the price of the electronic copy justified by its use? Do usage patterns show that some journals will be as heavily used—or more so—in 20 years as when they are published? Answers to these and other questions require statistics on usage, and in the electronic realm, such statistics must come from the publishers.

Unfortunately, it has been difficult, if not impossible, for librarians to obtain meaningful usage data from publishers of electronic journals. The reason is not a simple matter of publishers being unwilling to provide such information, even though some complain that implementing a data collection function is costly and others fear that librarians will cancel subscriptions if they learn that usage is low. A more basic problem is that there is no agreement on how to produce data that can be compared and analyzed. It has been exceedingly difficult for librarians to know what to ask for when something as basic as the term "use" can have many meanings.

CLIR commissioned Judy Luther to review how and what statistics are currently collected and to identify the issues that must be resolved before librarians and publishers feel comfortable with the data and confident in using them. In her extensive interviews with librarians and publishers, the author found significant common ground on the types of concerns held.

We hope that, in identifying some of the critical issues, this white paper will provide a basis for discussion among publishers, librarians, and aggregators that will lead to effective cooperation in collecting and analyzing usage statistics. CLIR will continue to pursue the agendas that librarians and publishers share as they make the transition to the digital environment and they find new ways of meeting users' needs.

Deanna Marcum
President

Introduction

Expenditures on journals comprise about 70 percent of the average academic library's budget for materials. Research libraries in the United States spend more than $500 million on journals annually, and several large libraries estimate that they spend 20 percent of their budgets on electronic materials.

When scholarly journals are converted to electronic form, they are frequently offered as part of a database hosted by a publisher or an aggregator and made accessible through the Web. As users shift from using local print materials to using remote files, librarians seek to collect usage data that justify the library's investment in electronic resources. Because the library licenses, but does not own, the content, it must depend on the provider for usage data.

Providing usage data is a new role for publishers and aggregators—one that requires not only much learning but also a financial investment. While it appears that the data would be as useful to publishers as to librarians, publishers must first develop the capability to serve their own purposes and then provide additional analyses and support to present the data so that librarians can use them.

Less than half of the publishers who offer journals in electronic form today are able to provide statistics on the usage of these journals. What is available varies widely among publishers, and librarians are often unclear about what to ask for and how they will use the data. Guidelines for compiling statistics are just emerging and have not been widely adopted.

Publishers are concerned that the data they share with librarians lack context. If, in the absence of such a context, usage data seem low, the publishers fear that librarians may use such information as a basis for canceling subscriptions. As both librarians and publishers become more familiar with the current state of usage statistics, the focus of the conversation will shift to what needs to be done to ensure consistency and to provide a valid context for understanding the data. There have been rapid developments in this area in the six months since this study began, and the author is encouraged by recent discussions with publishers who were previously reluctant to

provide data to libraries and are now inquiring about what should be delivered.

This paper provides a snapshot of developments in the industry. It identifies issues that concern both publishers and librarians and suggests a context for further discussion between the providers and consumers of electronic journals.

Background

Since libraries that host electronic journal content locally face the same challenges in collecting usage statistics as do publishers, the author chose to interview librarians at OhioLINK, Los Alamos National Labs, and the Florida Center for Library Automation to determine how they provide such data to their consortia members. Publishers and providers were then interviewed to compare their approaches to collecting and presenting the data with those of the libraries. Finally, comments from both groups were solicited to identify their concerns and to establish a base for understanding and interpreting the data. Summaries of selected interviews are provided in Appendix A.

To provide a frame of reference for this study, current initiatives by other organizations on developing data collection policies were reviewed. In addition to guidelines published by the International Coalition of Library Consortia (ICOLC), which were based on the JSTOR initiative, there is substantial work being done in this country by the Association of Research Libraries (ARL), the National Information Standards Organization (NISO), and the National Commission on Libraries and Information Science (NCLIS), and abroad by the European Commission. The ICOLC guidelines may be found in Appendix B; information on related industry initiatives appears in Appendix C. While some of the studies focus on defining data elements and collecting statistics, others examine data in light of factors used to assess performance, such as the percentage of user population being served.

Issues Affecting Librarians and Publishers

Among the most important findings of this study is that librarians and publishers share a significant number of concerns about the development and interpretation of statistics. Both are seeking agreement on core data that are needed and are exploring an appropriate context for interpretation. Once publishers and providers discover how to produce comparable and reliable data, it will be possible to continue discussions about usage and value to the user.

Issues of Common Concern to Librarians and Publishers

All indicators of usage are steadily rising, in part because of the continued growth of electronic content available on the desktop. However, in the electronic world, there are more variables that affect the

analysis of statistics and an understanding of the results than there are in the print world. For a balanced picture, librarians and publishers will want to consider how the following variables affect their data, assessments, and conclusions.

Lack of comparable data

The issue of greatest concern to publishers and librarians is the lack of comparable data. Variations in definitions and implementation procedures make it impossible to compare data from different host environments with any degree of reliability.

Unless data on multiple publishers are collected from the same platform (such as OhioLINK, HighWire, or Catchword) with common hardware and shared software, variations in how items are identified and counted will skew the results. What is counted (e.g., searches, abstracts displayed, HTML pages viewed, PDF documents downloaded) and how (whether internal use, such as demonstrations, and external use, such as spider hits or "robot contamination," are excluded) will vary according to the software used.

Librarians currently receive reports with different data elements that are not clearly defined and that cover different time periods, making it impossible to analyze them in a consistent way. Publishers likewise find it difficult to reconcile internal data that are produced from different systems that count data in different ways.

Lack of context

With insufficient data from the print environment and insufficient experience in the rapidly changing electronic environment, it is not possible to establish a context for understanding data available on the current level of online activity. What little data librarians have on the use of print cannot serve as a basis for projections on the use of electronic journals.

Current measures are limited to data on the amount of activity, such as the number of downloads. To base comparisons on the use of large or very popular journals (such as *Nature*, *Science,* or *Cell*) sets an artificially high benchmark for other titles with fewer articles available for use. This raises the question of whether the measure of activity should be relative to another factor, such as the price of the journal or the number of available articles, which puts the measure in a context.

Both publishers and librarians emphasize that measures of the level of activity do not indicate the value of an article. It is dangerous to assume that a popular title that is used by many students is worth more than a research title that is used by only a few faculty members working in a specific discipline. Other factors need to be considered.

Known differences in information-seeking behavior among users in various scientific disciplines warrant additional study to identify usage patterns. As more data are examined on use and behavior, it may be possible to establish average levels of use for different subject areas or user groups.

Incomplete usage data

Constructing a complete picture of use is further complicated by the existence of journals in multiple formats that are available through multiple sources or distribution channels, e.g., directly from the publisher, onsite at a library, or through a vendor such as OCLC. This means publishers must combine usage data for journals that are mounted on a remote host, such as OCLC or OhioLINK, with data for journals kept on their own Web sites. Libraries are confronted with multiple sources of usage data, or the lack thereof, for different formats (print, electronic, microfilm) and for multiple copies of titles that are available from several sources.

Marketing

Publishers who make usage data available are aware that this information will be used to assess the value of their journals. Consequently, they want to ensure that usage is high so that the cost per-use is low compared with that of other publications. Publishers and librarians with experience in electronic databases agree that marketing to users—whether librarians or library patrons—and making them aware of the availability of the resource and its features have a noticeable impact on usage.

It can take from sixteen months to three years for users to integrate into their routines changes in how they access information. For that reason, the amount of time a database has been available influences usage rates (Townley and Murray 1999). Elsevier's experience with The University Licensing Program (TULIP) and Pricing Electronic Access for Knowledge (PEAK) taught publishers that it is essential to promote the availability of a journal database to users and to allow time for user behavior to change.

Both librarians and publishers involved in the PEAK project acknowledged that publicity and promotion made a difference in levels of use. At Vanderbilt University, the medical school's use of the electronic journals was disproportionately low because the medical library was reluctant to publicize the use of a system that its staff considered to be temporary (Haar 2000).

Content provided

The demand for specific electronic titles is affected by both the timeliness of content and the amount of content provided. Some publishers release articles in electronic form before publishing the print version or choose to delay the electronic version for a few issues or for a year so as not to affect current subscriptions.

A collection becomes more useful when the amount of archival content available online increases, especially if it is well indexed. When back files are included with the current subscription or basic service, the user has more articles to view, and this will affect usage.

Interface affecting usage

Barriers to access, such as requirements to register, have proved to be a major deterrent to use (Bishop 1998). Charging user fees also limits

access. Only in an unrestricted economic environment can demand be measured accurately.

The presence of links that take the user directly to the full text of articles from the library's online catalog or from a Web list of journal holdings results in higher usage levels. Access data from Elsevier and MCB indicate that a high percentage of current use reflects the behavior of researchers who browse by a familiar journal title rather than that of general users who are searching for information on a subject. Tom Peters, director of the Center for Library Initiatives at the Committee for Institutional Cooperation (CIC), believes that accessibility is one of the crucial and complex factors affecting use.

The user's experience of the interface also will significantly affect the results. Both Academic Press and the American Institute of Physics (AIP) noted that they experienced surges in usage after they introduced new platforms that simplified navigation and access.

Economic model

As long as the journal, rather than the article, is the primary unit of sale, statistics will be collated by journal title. Academic Press pioneered a site license for consortia that includes all journals published by the Press and gives the user unlimited access to all articles. In this model, titles that are highly used will have a lower cost per-use and be perceived as a better value.

The trend toward offering a large database of journals from which the user selects articles gives rise to new economic models. The PEAK project, in which Elsevier titles were loaded at the University of Michigan, allowed users to access, for a nominal additional cost to the library, articles from journals to which the university did not subscribe.

Some librarians have begun to develop analyses based on article usage and cost per article. The hazard of pricing per-use is that value is associated with productivity of an article rather than with other measures. Pricing solely by usage may work with popular titles, but it ignores the importance of little-used titles that have an impact on research.

User privacy

The topic of privacy applies both to data collected on individual users and to data on libraries shared within consortia.

Data that publishers currently provide on journal use do not reveal specifics about any individual user, but present a summary of activity by journal title. However, publishers who offer personalized or customized services, such as e-mail alerts, must retain user-specific information in order to deliver such services, and this requires that they establish policies about how they intend to use the data.

Librarians have a tradition of protecting the privacy of users with policies regarding book circulation records. They are equally committed to protecting users' rights in the electronic environment. Publishers are considered responsible for how they use data they collect. Protecting the user's personal information is not just a courtesy:

it is a legal obligation (Rothman 2000). Publishers that collect such data need to develop policies for how it will be used throughout their organizations. Moreover, after a policy is established, it is essential that the company monitor compliance internally.

While the U.S. Federal Trade Commission is concerned that companies adhere to the privacy policies that the companies themselves have defined, there are more stringent requirements in Europe. Publishers who sell electronic publications in Europe must have privacy policies that indicate what information is collected, how it is used, how the user can change it, with whom it is shared, and how users can opt out.

For example, Elsevier's ScienceDirect product alerts users to the fact that they may be entering personal information when they take advantage of customized services or order documents. The company's privacy policy states that "Your information is kept confidential, unaltered and is used only by ScienceDirect and its parent company Reed Elsevier to administer your ScienceDirect/Elsevier Science relationship."

JSTOR is an industry leader in the area of statistics. Its privacy policy states that "No data are provided that would allow for the identification of the activity of individual users. Librarians can generate reports only for their own institution's usage activity, although average usage activity at similar institutions is provided for comparison purposes." The concern for privacy at some libraries extends to JSTOR's own statistics, and the library may want its usage data shared only with its permission.

In its guidelines, ICOLC states that "Statistical reports or data that reveal confidential information about individual users must not be released or sold by information providers without permission of the consortium and its member libraries" (1998).

Library Issues

While usage statistics validate the library's investment, they also provide insights into usage patterns that indicate the need to access a broader spectrum of titles than previously owned. This raises questions about the approach to building collections on a "just-in-case" basis compared with new models that incorporate on-demand acquisition.

Budget justification

Reference librarians lament that students act as if a resource does not exist if it is not online. Declining book circulation and rapid growth in the use of electronic resources indicate that users are shifting from print to electronic resources.

Libraries can tell which Web sites users are going to for information, but once users reach the publisher's site, their activity can be tracked only by the publisher. This means the library is dependent on the publisher to provide it with data vital for its internal reports. High usage demonstrates a good investment to administrators who

approve budget increases. For example, one library used statistics on after-hour use to show how the availability of electronic journals extended the library's services.

Impact on selection

Recent data from OhioLINK show that more than half of the articles selected by users come from journals not currently held by the library (Sanville 2000). There is increasing evidence from both libraries and publishers that current holdings are too limited to meet user demand, a trend that points to the benefits of user-driven selection procedures. The emerging models for article selection from a database of electronic journals challenge libraries to restructure their approaches to collection development and create new models to meet their users' needs.

Publisher Issues

Publishers who have experience with their own usage statistics are becoming less worried about cancellations because they see that librarians are still processing the data, rather than reacting to it. Many publishers are still concerned, however, that because there is no context for most usage data, it can be misunderstood.

Internal applications

As publishers come to terms with the costs of developing their capability to collect and analyze usage statistics, they find multiple applications for such information internally. For example, the systems staff uses such data to budget for new hardware. The product-development staff analyzes how users access content. Marketing is interested in how users find the site. The sales staff wants to know about the level of activity of their customers, and the editorial staff wants data on the most requested search terms.

Quantitative Measures

The establishment of accepted means for producing reliable and useful data can be viewed as a two-phase process. In phase one, publishers reach agreement with each other and with librarians about what data are required and what standards should be adopted for collection and delivery. Once comparable data are available, it will be possible to analyze and draw conclusions from the data—phase two of the process.

ICOLC guidelines were created to address a variety of files, including bibliographic databases, which are focused on simultaneous users. As a result, they reference turnaways and menu selections that do not apply to sitewide licenses for access to journal databases.

There is the potential to learn a great deal about users and their behavior; however, at this early stage, experienced librarians agree that it is best to focus on only a few measures. OhioLINK Executive Director Tom Sanville notes that of the criteria in the ICOLC guide-

lines, the only applicable measure of use of an electronic journal is the number of times an article is viewed, printed, e-mailed, or downloaded.

There are three steps in processing the raw data that servers produce on visits to the sites of Web-based journal collections.

1. *Data Collection*: On the basis of the needs of internal and external users, each host site decides what data elements it will collect. For example, do downloads include both HTML pages viewed on the screen and PDFs downloaded? Sizable log files are reviewed to extract and summarize data. Rather than use locally developed software, systems staff often prefer commercially developed software because it usually offers more features, enhanced graphics, and customer support.
2. *Analysis*: Decisions are made as to whether the analysis will be performed ahead of time on preselected fields or whether librarians can select the data elements, specify a time frame, and create their own reports.
3. *Presentation*: Once the content of the report is determined, decisions must be made on the currency of the data (i.e., real-time or periodic uploads on a nightly, weekly, or monthly basis), whether the files can be exported, and whether the data are pushed to the library via e-mail or the library must retrieve them.

What Are We Measuring?

Once publishers agree upon basic data elements to be collected, analyzed, and presented in a standard way, they will be able to produce the first generation of comparable statistics. Typically, what is being used (content), who is using it (user), and how the database is being used (activity) are measured. When the content is used and how the data will be presented are other questions of interest.

What is being used?

For a full-text journal database, the ICOLC guidelines define the use of articles as viewing, downloading, printing, or e-mailing the full text. Summaries of data usage by journal title can help librarians decide what titles to add, change, or delete and can assist publishers in determining the health of the journal. Comparing statistics on the abstracts and tables of contents viewed with statistics on downloads may provide insights on how users navigate the database.

Who is using the content?

Analysis by IP address range can sometimes reveal the academic department that has requested the article and can be useful in assessing the need to train users or make them aware that a resource is available. When users are remote and are assigned a dynamic IP address, it is difficult to determine the user's discipline. Some libraries attempt to combine usage data provided by the publisher with their

own data to determine the extent to which they are serving remote users.

How is the database used?

A "hit" registers each time the server receives a request to act (e.g., to do a search, to view an abstract, or to download an article). The type of hit can indicate how the user approaches the system—for example, to browse a journal title or to search for specific information. Searches include requests to search by title, author, or subject. Browsing includes accessing the full text from the journal title, issue, and table of contents. In menu-driven systems, the menu items that the user selects need to be counted. Direct access with a citation or URL may be counted separately.

When is the content being used?

Measures of activities (hits, sessions, downloads) are summarized by the hour, day, week, month, and year. The systems staff analyzes data from server logs to determine the ability of the server to meet the load during periods of peak demand. When systems provide access to simultaneous users, the number of times users are turned away also needs to be captured to measure unmet need.

How will the data be presented?

The degree to which statistics are useful to a library depends on how the data are presented. Librarians want to be able to do the following:

- Query the system and specify the time period covered
- Access two years of data online to monitor growth
- Download data as a "comma delimited" file to load into a spreadsheet
- Graph usage across years or titles or compare usage with that of other libraries
- Access data that are real-time or that are updated nightly
- Establish a profile and routinely receive the results by e-mail

Once publishers have established a well-defined and consistent set of data, additional analysis will support exploration of data related to behavior and use and will attempt to address questions related to value.

Data Reliability

A standard methodology for collecting and analyzing data is necessary to ensure that both publishers and librarians have data that are comparable and reliable.

With a full-text journal database, the conversation centers on three measures: hits (equated to searches), sessions (equated to users), and documents used (equated to downloads). However, measuring hits or sessions can yield misleading information. The number of hits will vary, depending on network access and telecommunication

factors. Likewise, the number of sessions will vary because of timeouts and other network protocols.

Conversations with the staff who implement the statistics function revealed a common process of learning related to the design and development of internal processes to produce valid data. Any given method of implementation can produce varied results, based on the software selected and the diverse nature of local systems architecture.

Caching: Caching allows frequently accessed Web pages to be stored on a server to improve performance. When users access cached materials, these actions are not counted as a hit on the host database. Consequently, for popular materials, statistics supplied by the host are likely to underestimate usage.

Log files: Web server log files are a good means of helping administrators gauge the demands on a server. Such logs measure requests for specific documents on a server, but they cannot show exact usage because caching is often employed and because users are assigned variable IP addresses (Bauer 2000).

Although log files are not designed to describe how people use a site, they do allow analysis of the source of links into a Web site and therefore can be used to determine which sites are generating traffic. Such information can be useful to publishers. For example, *Science News Online* (SNO) learned that one of its articles, which had been cited on another Web site that was linked to the SNO Web site, had brought in a high volume of visitors to the SNO site. As a result, SNO decided to mount the full text of the cited article (Peterson 2000).

Software: Bridget Pairaudeau, who handles the statistics function for Institute of Physics Publishing, noted that an off-the-shelf package such as NetTracker can be used to screen out robot contamination as well as data from internal testing, demonstrations, training sessions, and trials that skew the usage data. It was found that NetTracker records HTML articles viewed rather than PDFs downloaded; this could be a concern for publishers that offer both functions.

Summary and Next Steps

The evolution of electronic journals to publisher-hosted databases of journal articles shifts the burden of measuring use from libraries to publishers. Although publishers need to collect data for their own purposes, the associated costs are considerable. Additional work is required to produce data that are meaningful and useful for librarians.

Fears of conflicting motivations between publishers and librarians are diminishing as publishers become familiar with their own data and focus on the challenge of producing useful statistics. Concerns about comparability are valid and need to be addressed in a meeting where publishers who have already implemented statistical functionality can share what they have learned. Issues to be discussed at such a meeting would include producing useful data and interpreting the data.

Producing useful data

Publishers and librarians share concerns about the lack of standards for collecting and presenting data and the lack of context for drawing conclusions. Putting together a complete picture of use with data in multiple formats and from multiple sources is an additional challenge for both.

Given the variety of platforms and software packages, publishers need to learn from each other about the variables and to agree upon an approach that will produce consistent measures of use. If a group of individuals involved in producing statistics were to pool their intelligence and produce guidelines, it would greatly advance the state of the art. Once a critical mass of publishers is producing consistent data, it will be clear to others who are just beginning their work what data are needed and how to collect and deliver valid data.

The industry is at the first stage of creating the capability to gather statistics, establish standards, and deliver comparable and reliable data. In another year, new systems will emerge that rely upon data mining and analysis and that focus on understanding user behavior.

Interpreting the data

Users want and need access to a much broader range of material than that which can be owned affordably in print. Emerging pricing models and consortial arrangements that provide users with access result in data that show higher levels of use of nonowned titles. Interpretations of these data vary; for example, they include concerns that the right titles are not being bought as well as the recognition that the information industry is moving from a supply-driven model, with preselected packages of information, to a demand-driven model, where users choose what they need from a wide array of options. Making users aware of what is available and increasing the ease of access will require cooperation between publishers and librarians.

Librarians and publishers need to understand users and their information-seeking behaviors in ways that were not previously possible or necessary. As intermediaries between the author and the reader, publishers and librarians must learn how best to serve their users. Doing so will require further analysis.

Recommendations for next steps

Publishers are discovering what data they need to provide and how to provide it. There is no forum where staff working on statistics can share their understanding of the technology and make it easier for those who have just begun to tackle these issues. To facilitate the development of statistics in the industry, an organization such as the Council on Library and Information Resources might wish to sponsor an invitational meeting that would enable those involved in this area to discuss the issues they have encountered and to explore the development of guidelines for all participants.

Associations involved in creating standards and guidelines on data collection are focused on defining the data elements and determining what is currently being done. No one is working directly with publishers who have developed data, understand the variables, and are in a position to provide guidance so that those producing data can be consistent in their implementation. CLIR is well positioned to host such a meeting, which should include representatives from the publisher, vendor, and library communities. Preliminary feedback from publishers and aggregators has been favorable. The author welcomes additional input on the structure and desired results of such a forum.

APPENDIX A:

Interviews with Librarians and Publishers

OhioLINK, Los Alamos National Labs (LANL), and the Florida Center for Library Automation (FCLA) all host journal databases. They were selected for inclusion in this white paper because they had to develop the same capabilities being requested of publishers. Villanova University was included because it has closed stacks for its bound journals, which means that it has good measures of use. James Mullins, the university librarian at Villanova, was on the task force that created guidelines for the statistics that JSTOR delivers.

Academic Press, Elsevier, MCB, and the Institute of Physics (IOP) host their own journals and have experience with collecting statistics. The American Institute of Physics (AIP) and Association for Computing Machinery (ACM) are in the process of developing this capability.

JSTOR and Catchword both host content from a variety of publishers. JSTOR was part of the initial discussions about library requirements, while Catchword is further developing its statistics capability. Like the library hosts, these providers have a standard platform that provides consistent data to enable comparisons.

Libraries

OhioLINK

Because OhioLINK staff developed the statistics capability when they designed the overall system, its initial set-up costs are not readily identifiable. Ongoing support is provided by two staff members who perform many other duties.

In addition to issuing regular usage reports, OhioLINK has taken advantage of the opportunity to perform further assessment of usage. According to Executive Director Tom Sanville, this assessment shows that, although every title in the database has been used, 40 percent of the titles represent 80 percent of the downloaded articles, while another 40 percent of the titles received only 10 percent of the use (Sanville 2000). This prompted David Kohl, director of the library at the University of Cincinnati and a member of OhioLINK, to suggest that low usage might make the latter titles candidates for lower pricing (Kohl 2000).

A surprising discovery is that more than half (58 percent) of the articles downloaded for all OhioLINK libraries were not held in print by the libraries (Sanville 2000). In each institution, patrons make use of a much wider number of journals than those held in print.

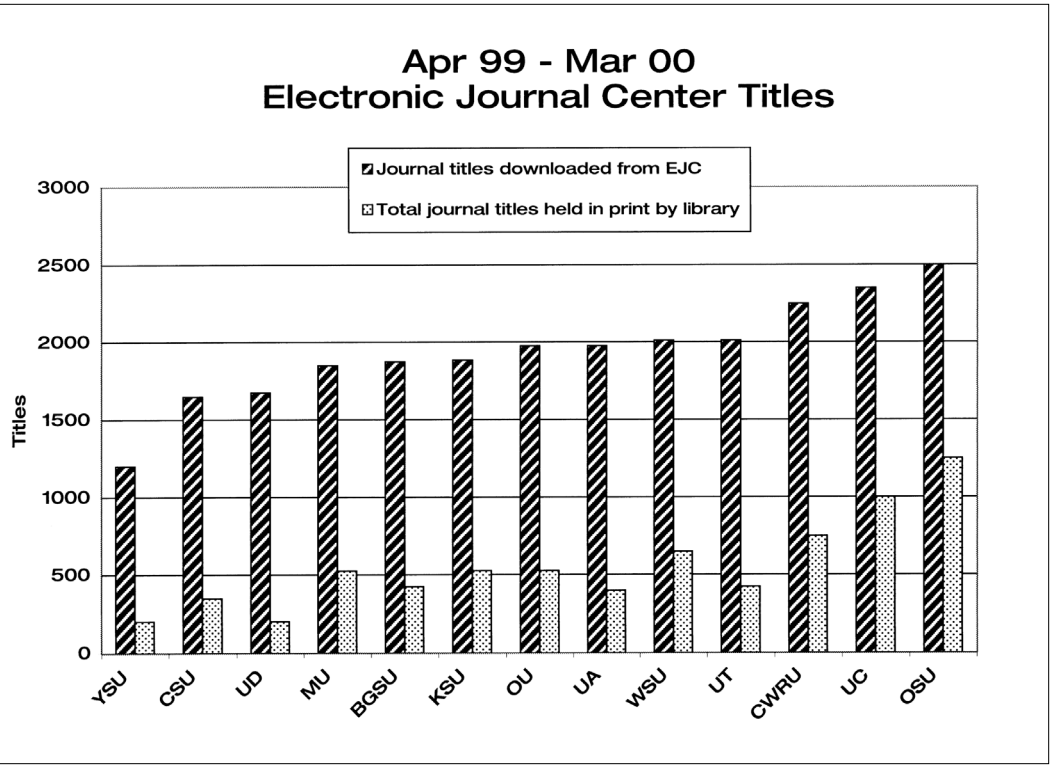
This finding speaks to limitations imposed by budgets on the selection process and the importance of letting the user choose from

a larger file of material. In a paper presented at Oxford 2000, David Kohl noted that presenting users with a database of journal articles allows them to drive the selection process in a way that is similar to the current practice under which vendors supply librarians with books on approval.

Los Alamos National Labs

The LANL Library has gone through three stages of development, according to Director Rick Luce. The data the library collects depend on how far it parses its log files. The first stage, which entailed parsing UNIX logs and "beat code," cost $20,000 and required nominal staff support. The second stage, which involved producing static usage data on the basis of scripted code, cost $50,000; one staff member performed this activity. The third stage, designed to enable the user to perform a query and export the results, may cost $250,000. Programming staff will be involved in doing the analysis.

LANL has 3,500 electronic journals available to its users; of these, 2,000 titles are loaded locally and 1,500 are accessed remotely. When LANL did a trial with Elsevier, all titles in the database were used and the participating libraries did not own the most-used titles.



BGSU - Bowling Green State University
CSU - Cleveland State University
CWRU - Case Western Reserve University
KSU - Kent State University
MU - Miami University
OU - Ohio University (Athens)
OSU - Ohio State University (Columbus)
UA - University of Akron
UC - University of Cincinnati
UD - Univesity of Dayton
UT - University of Toledo
YSU - Youngstown State University
WSU - Wright State University

Luce concludes that librarians do not know exactly what users need, confirming the discovery process in research and the learning curve in the electronic environment.

LANL enables its users to connect to full text from links within secondary publications, from browsing selected titles, and from performing subject searches. It takes six months for users to discover, remember, and fully use a new service. Keys to success are to ensure that links are established, to allow sufficient ramp-up time, and to promote awareness. LANL has expanded its electronic holdings since 1995, and user satisfaction with library services has increased dramatically.

Florida Center for Library Automation

FCLA is the central agency that supports the online catalogs of the 10 universities in Florida. Like OhioLINK and LANL, FCLA loads a number of full-text journal databases, for which it produces statistics locally as well as links to publishers' remote sites.

FCLA would like to track the number of searches, the number of documents retrieved, and the number of requests denied. The number of hits is not a valid indicator of use because there is no consistent way to measure them. The number of articles viewed by journal title is counted when the PDF is viewed. Reports on usage of full-text journals are updated nightly in a formatted report that the librarians can download.

When users link to a publisher's database, they have effectively left their home system. The library can tell which database they linked to, but it cannot track actions taken on the publisher's Web site. Consequently, libraries must rely on publishers for usage data and then merge such information with their own local data.

Villanova University

Villanova University Library Director James Mullins noted that students today rely solely on electronic publications because of their ease of access and use; consequently, they have a limited view of the available content.

Villanova can track the usage of its bound print journals because they are in closed stacks. Use of print journal collections was growing until 1995, when electronic databases were made available to users, who also began to access the Web. Since then, the library has seen a dramatic decline in the use of print materials and a steady increase in the use of electronic resources.

In an attempt to collect some data locally on student and faculty use of remote databases, Villanova analyzed its log summaries, which show the total number of times a database is accessed. These data are put into a spreadsheet as a frame of reference along with vendor-supplied data and are compared with the prior year's totals. Assistant Director for Public Services Louise Green noted that training usage should be counted separately so as not to skew the totals.

Publishers

Elsevier

Elsevier has at least two staff devoted to managing usage data from its ScienceDirect database installations. Most libraries subscribe to only a portion of the 1,170 titles in Elsevier's database; therefore, data on the use of nonsubscribed titles are helpful in considering the addition of electronic or print versions of a title.

Although Elsevier is committed to providing as much information as the customer believes is useful, staff acknowledge that custom reports are not economical to generate. The company can see the impact of marketing on journal usage, and it has a staff of account-development managers devoted to training librarians and users on the system. As the volume of articles used rises, the cost per use drops.

To keep current in their field, researchers scan about a dozen journals regularly by browsing their tables of contents. This activity is reflected in how the database is used when researchers select a journal title from a list and then browse the tables of contents of various issues, rather than search by subject, author, or title.

Elsevier has paid particular attention to global requirements for a privacy policy, which appears on a full page on the Web site for ScienceDirect. Some customized services, such as an e-mail address for an alerting service, cannot be provided if the user does not provide a minimal amount of personal information. To ensure privacy, all data on individual users are scrubbed at the organizational level before being processed and aggregated.

Academic Press

Academic Press found that the off-the-shelf software packages that summarize hits do not provide the data that libraries need. It is hiring a full-time statistician and measurement analyst to help address the issue. The company experienced a dramatic increase in usage when it introduced its new platform in the fall of 1999.

Data gathering is complicated because Academic Press's journal database (IDEAL) is loaded on remote sites such as OhioLINK and OCLC, and Academic Press needs to combine data from several sources for a complete picture of usage of its own journals. Data are used internally by sales, accounting, and editorial staff to examine correlations and draw conclusions about the cost per-article for each institution. This allows the publisher to understand how the library might equate the cost per-article to a relevant measure indicating value.

In the print world, subscription revenues indicate the health of a journal. When that journal is part of a database, the equation changes completely, since some of the articles used were in previously non-subscribed titles.

For every 1.5 log-ins to the database, one article is downloaded, and for every abstract viewed, there is one article downloaded. Academic Press summarizes the total number of log-ins by journal and

of articles downloaded by journal each month for each institution and consortium.

Chrysanne Lowe, director of online sales and marketing, noted that the journals that have the most articles downloaded are considered the company's most successful titles. These are large journals with many articles. The list of journals in greatest demand changes when the number of articles downloaded is compared with the number of articles published in the title.

Philosophically, Academic Press is opposed to a business model in which charges increase with use because it discourages use. Academic Press offers marketing support with promotional items and coordinates training with librarians and faculty members.

MCB University Press

In addition to the normal data on time-of-day activity that help it determine the load on systems, the system at MCB University Press tracks hits and sessions. To learn how users come to the site, MCB also analyzes the top referring sites, top browsers, top entry pages, and the most popular and least popular pages in the database.

MCB University Press is interested in knowing which institutions generate the most requests and which articles and journals are most requested. How users search is also of interest; for that reason, data on the tables of contents, search pages, and browse pages are collected.

Heavy use of the tables of contents through the browse functions indicates that many users know the title they wish to see. However, MCB discovered that the most-used titles at some institutions were the first titles in the alphabet. This indicates that users are learning how to use a system and suggests the need to evaluate the interface or provide more training.

Institute of Physics

Bridget Pairaudeau, producer of electronic publications at IOP, just completed the design of IOP's statistics form for internal use. It allows staff to select the following variables:

- *Who*: user files and the subscription records from IOP's internal systems
- *What*: data from log files on the type of activity and time frame
- *View*: display options, such as grouping subscribed journals

Users of the IOP system also have the option of creating a graph by selecting elements for the *x* and *y* axes. If they chose to graph usage of Web pages on both axes, they can show navigation to full text from the table of contents compared with navigation from the subject keyword search. Data on the use of options that can be customized, such as profiling, use of filing cabinets, and activating a table of contents alerting service, show which features are most used.

The editorial and marketing staffs are interested in knowing which articles and journals are most requested and which institutions are most active. The sales department is interested in the level

of use by specific customers, and system designers want information they can use to enhance features, navigation, and usability.

IOP screens out data on internal use, guests, free use, trials, production applications, and robot attacks, because they can greatly skew statistics. When IOP's internal data analysis did not match that of the commercial package, staff discovered that NetTracker counts HTML views but not PDF downloads.

American Institute of Physics

Doug LaFrenier, director of marketing at AIP, noted that the market has changed dramatically. Providing statistical data to libraries represents a new set of responsibilities for publishers—one that has associated costs. LaFrenier's primary concern is the lack of standards, which makes it impossible to compare data.

AIP is concerned that it is undercounting because its system does not count searches and requests for abstracts. It counts only requests for the full text of an article that requires either a subscription or pay-per-view access. At the same time, AIP has discovered that one of the interfaces was triple counting downloads because of the way it grabbed the content.

The American Institute of Physics, working with the American Physical Society (APS) has devoted much of one full-time programmer's activity to developing Web-based statistics that libraries can access for their own use. The statistics, which will be available to other publishers that AIP hosts, are planned for delivery early in 2001.

AIP demonstrated the system at the Special Libraries Association 2000 meeting. The demonstration showed year-to-date download statistics. Libraries who attended this session persuaded AIP that libraries want to be able to specify their own time periods. They also want to be able to compare current data with information from prior years. AIP found it difficult to identify who within the library should have rights to view this information.

Previously, AIP had given its own publishing customers reports from the server logs that summarize activity by journal title. The company also has analyzed time-of-day performance data to support decisions in running an online journal platform. It has been able to identify the most active journals and accounts, and believes that much of the information developed for online publishing customers will be useful in developing usage-statistics reports for libraries.

Anyone using the AIP Web site has the option of buying an article online. Sales grew significantly when the company simplified its interface and reduced the number of steps required for the user to obtain the article. This further supports the importance of ease of interface on usage.

Association for Computing Machinery

The Association for Computing Machinery is evaluating what statistics need to be collected. As staff experimented internally with data, they found that the most frequently downloaded article in any given month was neither a current article nor one they would have expect-

ed to be so popular. High-use article titles provide clues for editors about the topics in demand.

Providers

JSTOR

The ICOLC guidelines are based on those developed by a task force in conjunction with JSTOR in 1997. JSTOR data are updated nightly and can be queried and exported to a spreadsheet. Individual site data can be compared with average data for all sites in the same JSTOR classification and with summary data for all JSTOR titles. Both publishers and librarians can sign on and retrieve data.

Data presented include the number of pages viewed, PDFs printed, searches conducted, and tables of contents browsed. Since JSTOR includes as articles all items (e.g., reviews and letters), it lists full-length articles separately for clarity.

In a presentation at the Conference on Economics and the Usage of Digital Library Collections, JSTOR President Kevin Guthrie observed that the articles that are most often downloaded are not those that advance research or that are most often cited (Guthrie 2000). "Value needs to be clearly defined as libraries consider acquisition and cancellation decisions for electronic content," Guthrie stated. (Marthyn Borghuis from Elsevier noted that citations reflect author activity while usage reflects reader activity).

The notion of perishability of content varies with the discipline. The average age of the most-used articles was also surprising: 13 years in economics and 32 years in mathematics. When there are a small number of total accesses for the discipline, the actions of a few people can sway the results.

Guthrie cautioned that usage does not necessarily equate to value in the research sense. "Older articles may be absolutely vital to the continuation of high-quality scholarship and research in the field, but that may not lead to extensive use," he said.

Catchword

Catchword delivers service that is paid for by the publishers, who decide what information to share with libraries. Catchword is expanding its statistics ability according to ICOLC guidelines, and it will have data that can be used by libraries. Catchword has decided to add turnaway statistics that reflect the number of times a user attempts to access the full text of an article in a journal to which the library does not subscribe. Catchword can also track pay-per-view access. Although the company has a single source to produce these data, its challenge is to summarize data from 11 servers around the world.

HighWire Press

HighWire Press has developed extensive data analysis and reporting capabilities for publishers and librarians who can download their report from HighWire to Excel. Journal usage data includes: statistics

on the volume of searches, table of contents, abstracts, articles viewed in HTML, and PDFs downloaded. Demand for articles is measured by: the number of unique articles and total accesses by abstract, HTML views, and PDFs downloaded. It is also possible to see the top ten articles in each journal ranked by total accesses (HTML, PDF, abstract) with an indication of the age of the article. As part of a Mellon- funded grant to Stanford University Libraries, HighWire transaction logs will be analyzed using data mining techniques to uncover user behavior and trends.

APPENDIX B:

ICOLC Guidelines

Statistical Measures of Usage of Web-Based Indexed, Abstracted, and Full Text Resources (November 1998)

1. Requirements

Each use element defined below should be able to be delineated by the following subdivisions:

- by specific database provider
- by each institutionally defined set of IP addresses/locators to subnet level
- by total consortium
- by special data element passed by subscriber (e.g., account or ID number)
- by time period. Vendor's system should minimally report by month. For each month, each type of use should be reported by hour of the day, and vendor should maintain 24 months of historical data.

Use elements that must be provided are:

- Number of queries/searches categorized as appropriate for the vendor's information. A search is intended to represent a unique intellectual inquiry. Typically a search is recorded each time a search form is sent/submitted to the server. Subsequent activities to review or browse among the records retrieved or the process of isolating the correct single item desired do not represent additional searches, unless the parameter(s) defining the retrieval set is modified through resubmission of the search form, a combination of previous search set, or some other similar technique.
- Number of menu selections categorized as appropriate to the vendor's system. If display of data is accomplished by browsing (use of menus), this measure must be provided (e.g., an electronic journal site provides alphabetic and subject-based menu options in addition to a search form. The number of searches and the number of alphabetic and subject menu selections should be tracked).
- Number of sessions (logins), if relevant, must be provided as a measure of simultaneous use. It is not a substitute for either query or menu selection counts.
- Number of turnaways, if relevant, as a contract limit (e.g., requests exceed simultaneous user limit).
- Number of items examined (viewed, marked or selected, downloaded, emailed, printed) to the extent these can be recorded and controlled by the server rather than the browser.

- Citations displayed for A&I databases
- Full text displayed by title, ISSN with title listed or other title identifier as appropriate.
  1. Tables of Contents displayed
  2. Abstracts displayed
  3. Articles or essays, poems, chapters, etc., as appropriate viewed (e.g., ASCII or HTML) or downloaded (e.g., PDF, email)
  4. Other (e.g., image/AV files, ads, reviews, etc., as appropriate)

2. Privacy and user confidentiality

Statistical reports or data that reveal confidential information about individual users must not be released or sold by information providers without permission of the consortium and its member libraries.

3. Institutional or consortial confidentiality

Providers do not have the right to release or sell statistical usage information about specific institutions or the consortium without permission, except to the consortium administrators and member libraries. Use of institutional or consortium data as part of an aggregate grouping of similar institutions for purposes of comparison does not require prior permission as long as specific institutions or consortia are not identifiable. When required by contractual agreements, information providers may furnish institutional use data to the content publishers.

4. Comparative statistics

Information providers should provide comparative statistics that give consortia a context in which to analyze statistics at the aggregate institutional (consortium member) level. For example, a grouping for purposes of comparison should be compiled by the information provider (e.g., statistics from an anonymous selection of similar institutions), or it might be a grouping composed on demand (e.g., statistics from all campuses in a consortium, presented either anonymously or not, as desired by the participating institutions).

5. Access / Delivery mechanisms / Report formats

Access to statistical reports should be provided via web-based reporting systems and be restricted by IP address or another form of security such as passwords. Institutions should be able to authorize access to their data by other institutions in the consortium if they desire. Information providers should maintain access to tabular statistical data through their web site (updated monthly) which a participant can access, aggregate and manipulate on demand. When appropriate, these data also should be available in flat files containing specified data elements that can be downloaded and manipulated locally. Information providers are also encouraged to present data as graphs and charts.

APPENDIX C:

Related Industry Initiatives

National Commission on Libraries and Information Science (NCLIS)

http://www.nclis.gov/libraries/lsp/statist.html

Denise Davis at the NCLIS has commissioned John Bertot and Charles McClure of Florida State University to undertake a project entitled the "2000 Internet Connectivity Study." The authors will measure the level of connectivity, public access, training support, and technology available for the staff and patrons of public libraries. Focusing primarily on aggregators of indexes that include full text, the study authors are gathering information on the ability of public libraries to report electronic database use.

Association of Research Libraries (ARL)

http://www.arl.org/stats/newmeas/newmeas.html

Martha Kyrillidou manages ARL's New Measures Initiative, which includes E-Metrics, a major project that began in June 2000. E-Metrics focuses on the development of statistics and performance measures for the delivery of networked information resources and services. Twenty-three ARL member libraries are participating in a study led by Charles McClure and Wonsik (Jeff) Shim from the Information Management Use and Policy Institute at Florida State University.

Scheduled for completion in December 2001, the E-metrics project has three phases. In the first phase, information will be gathered on ARL libraries' best practices in statistics, measures, processes, and activities that pertain to networked resources and services. In the second stage, a methodology will be developed to assess the degree to which such data collection is possible and collected data are comparable among member libraries. In the third phase, a set of refined measures with data descriptions and guidelines for data collection, analysis, and use will be proposed. A separate task force within the project will focus on vendors' statistics, i.e., the definition of data elements and terms, specific data that can be collected, and methods for reporting data to libraries.

National Information Standards Organization (NISO)

http://www.niso.org/

Patricia Wand, director of the library at American University, and Denise Davis, director of statistics and surveys at the NCLIS, are leading the planning process for a review of the current standard on Library Statistics. This revision will address areas such as performance measures and the measurement of electronic services and re-

sources, which were not dealt with in the last review. Formal discussions will begin at a workshop to be held in February 2001.

EQUINOX

http://equinox.dcu.ie./reports/pilist.html

Funded by the European Commission, EQUINOX is designed to gain agreement on performance measures for the electronic library and develop an integrated software tool for use by European librarians. Building on earlier projects that focused on tools for book collections (e.g., EQLIPSE, MINSTREL), EQUINOX will take the lead in developing electronic performance indicators.

The indicators in this project are defined either in an International Standards Organization (ISO) document or by the project team. Team members use several methods to identify the percentage of a library's target population that is served and trained to use materials. These include on-site and remote sessions, downloads, cost per session and per download, level of workstation usage and number of rejected sessions, and percentage of the acquisitions budget spent on electronic resources.

LibEcon 2000

http://www.libecon2000.org

Funded by the Directorate General X (DG X) of the European Commission, this three-year project is nearing completion. It is focused on gathering consistent information about the libraries' development as information resources within European countries. The LibEcon 2000 Web site was established to test and then generate an automatic means of collecting data from respondents in 29 countries.

Project staff are working closely with the United Nations' Educational, Scientific, and Cultural Organization (UNESCO); the International Federation of Library Associations and Institutions (IFLA); the European Commission's central statistical agent (Eurostat); the European Bureau of Library, Information, and Documentation Associations (EBLIDA); and the appropriate committees of the ISO.

REFERENCES

Association of Research Libraries. 2000. New Measures: Developing Statistics and Performance Measures to Describe Networked Information Services and Resource for ARL Libraries: Discussion Prospectus. Washington, D.C.: Association of Research Libraries.

Bauer, Kathleen. 2000. Who Goes There? Measuring Library Web Site Usage. *Online*. (January). Available from http://www.onlineinc.com/onlinemag/OL2000/bauer1.html.

Bertot, John. 2000. Developing National Library Network Statistics & Performance Measures. Available from http://www.albany.edu/~imlsstat.

Bishop, Ann Peterson. 1998. Logins and Bailouts: Measuring Access, Use, and Success in Digital Libraries. *The Journal of Electronic Publishing* 4(2). Available from http://www.press.umich.edu/jep/04-02/index.html.

Guthrie, Kevin. 2000. Revitalizing Older Published Literature: Preliminary Lessons from the Use of JSTOR. Paper presented at the Economics and Usage of Digital Library Collections, Ann Arbor, Michigan, March 23–24.

Haar, John. 2000. Project PEAK: Vanderbilt's Experience with Articles on Demand. *From Carnegie to Internet2: Forging the Serials Future, Proceedings of the North American Serials Interest Group, 14th Annual Conference (Carnegie Mellon University, Pittsburgh, Pennsylvania, June 10–13, 1999)*. Binghamton, N.Y.: Haworth Information Press.

International Coalition of Library Consortia (ICOLC). 1998. Guidelines for Statistical Measures of Usage of Web-Based Indexed, Abstracted, and Full Text Resources. Available from http://www.library.yale.edu/consortia/webstats.html.

Kohl, David. 2000. Shifting Approaches to Collection Development: Should We Bother Selecting Journals at All? Paper presented at Oxford 2000: The Fiesole Collection Development Retreat Series, Oxford, England, July 20–22.

Peterson, Ivars. 2000. Beyond Hits and Page Views. *The Journal of Electronic Publishing* 5(3). Available from http://www.press.umich.edu/jep/05-03/peterson.html.

Public Library Association. 2000. PLA Tech Note: Electronic Statistics: Counting Crows. Available from http://www.pla.org/technotes/electronicstats.html.

Rothman, Joel B. 2000. Establish an Effective Privacy Policy. *e-Business Advisor (*March): 34.

Sanville, Tom. 2000. A Method Out of the Madness: OhioLINK's Collaborative Response to the Serials Crisis, Three Years Later–Progress Report. Paper presented at the North American Serials Interest Group Conference, University of California, San Diego, June 22–25.

Townley, Charles, and Leigh Murray. 1999. Use-Based Criteria for Selecting and Retaining Electronic Information: A Case Study. *Information Technology and Libraries* 18(1):32–9.