online information 2007
Day 2 – Track 3

What does usage data tell us about our users?

Carol Tenopir, Eleanor Read, Maribeth Manoff and Gayle Baker
University of Tennessee, USA

David Nicholas
CIBER, University College London, UK

Donald W. King
University of North Carolina, Chapel Hill, USA

Introduction

The more librarians and publishers know about how users access and read their materials, the more equipped they are to design better information products and services. Reading scholarly journal articles is especially important for academic staff, students, and other researchers to assist in all aspects of teaching, learning, and conducting new research. Many studies over the years have provided a longitudinal look at how electronic resources have influenced reading patterns (see for example, Tenopir and King, 2000, 2004.)

Usage data for electronic resources can assist with collection development, reveal which systems and sources are most often used, and show how users access our collections.

E-journal collections can generate log data on views and downloads, but log data do not always tell much about individual users. With deeper scrutiny, data can also reveal information about the users behind the usage data. Deep log analysis, library-collected usage data, and surveys of users together can reveal a picture of users or user groups in addition to the surface picture of usage of materials. This presentation explores how various types of log data and surveys can be used to show patterns of use by academic staff and students and the value and purpose of reading scholarly journals. The focus will be on a user-centred approach to answer questions about the faculty, staff, and students and how they use library collections, rather than a materials-centred approach that answers questions about the e-journals collection.

In a three-year project called 'Maximizing library investments in digital collections through better data gathering and analysis (MaxData)', sponsored by the US Institute of Museum and Library Services, we have gathered and compared various types of usage data from five US universities. (In addition, we have survey data from two Australian universities from a previous project.) Understanding users as well as usage is the ultimate goal. Data about our users can tell us many things about the academic staff, students, and researchers who need information, including what they use now, what they prefer to use (and what they do not like to use), and, by extrapolation, what they might use in the future. Comparing the various methods to help libraries determine which methods are appropriate for the questions they need answered is another goal.

Methods

Three main methods for collecting data about the users who access library journal collections are being compared in the MaxData project. Two of these focus only on e-journal usage. The first of these, deep log analysis, uses in-depth analysis of integrated transaction logs. The second, usage data analysis, examines the variety of vendor-provided (both COUNTER-compliant and non-compliant reports) and usage statistics collected by the library (such as those generated by link resolvers). The third method of gathering information about usage and users is by surveying library users. The surveys provide self-reported information about the use of both print and electronic journal articles from the library collections and other sources and include explicit measures of the value and outcome of reading journal articles.

The deep log analysis in this project was conducted by a team at University College London's Centre for Information and Behaviour Evaluation and Research (Ciber). The team analysed usage logs from the OhioLink consortium, a state-wide consortium in the US state of Ohio. At the time of the study (2005–6), OhioLink locally loaded and provided usage data for over 6,600 journal titles. This is an integrated e-journals system, where the local search and retrieval system provides access to journals from many different publishers to all of the members of the consortium. The usage information for the four Ohio universities for which we also have survey data can be isolated in the logs for a one-on-one comparison.

Library data analysis focused on one university's local e-journals collection. The University of Tennessee (UT) is a research extensive university with a student population of over 26,000 and more than 1,400 faculty members. The UT Libraries provide a collection of approximately three million items, including an extensive collection of electronic journals, e-books, and databases. Metasearch capabilities are available, as well as individual searching by database or e-collection title. Link resolver software provides links to full-text articles from multiple sources. Remote access to IP-authenticated resources is provided through a proxy server.

The UT and Ciber teams are both working with usage data, but from different sources and with different applications. UT's usage data are collected from vendors and standard library systems (for example, link resolver), and are used to inform collection management decisions for

electronic resources. The data are summaries of log data, not actual log data with detailed information, and are processed and analyzed with a combination of programming and manual manipulation. The Ciber team's data, on the other hand, are log data from the OhioLINK servers, and therefore are consistent and rich with information that can reveal much about users' searching styles and preferences, decay in article usage, and other things, in addition to how much various resources are being used.

Finally, surveys were conducted of faculty and students at four Ohio universities and the University of Tennessee. All survey instruments and individual reports can be found at http://web.utk.edu/~tenopir/research/. The ten survey instruments for all five universities were identical except for some variation in the demographic question regarding subject discipline of the respondent. For analysis this question was collapsed into five broad categories of subject discipline (science, engineering/technology, social science, humanities, and medicine/health). The surveys were available online for respondents in October and November 2005. A cover letter from the local university's director of libraries with an embedded link to the survey instrument was sent to a sample of faculty and students in October, with a follow-up at the end of the month. Over 1,300 faculty and 3,600 students in the five US schools responded to the questionnaires, with an additional 381 faculty and 863 students in two Australian universities in surveys conducted earlier in the year.

The strengths and weaknesses of each method and what each method reveals about users, rather than collections, are described here.

Methods: more about deep log analysis

Deep log analysis is a specific form of transactional log analysis, which tells us much about the information seeking behaviour of users but only a little about user satisfaction and information need. By definition, it also only provides information on users of a digital system, nothing on non-users of the system or hard-copy users. Generally the attractions of log analysis are:

- Logs record the use of everyone who happens to engage with the system and therefore the data yield and reach is absolutely enormous – half a million users in the case of OhioLINK. There is no need to take a sample and thus questions of representativeness do not arise.
- They are a direct and immediately available record of what people have done: not what they say they might, or would, do; not what they were prompted to say, not what they thought they did. Logs do not rely on memory.
- The data are collected routinely, automatically and quite anonymously. There is no need to contact the user or obtain their co-operation.
- Logs provide a level of detail not obtainable by any other method. They record everything that someone does online, while they are viewing, searching, browsing and navigating over the survey period.

Deep log analysis (DLA) differs from the traditional form of log analysis as found in the log reports libraries obtain from publishers/vendors in that it is based not on proprietary software but on an SPSS analysis of the raw server logs. This has some advantages.

Even greater detail can be furnished regarding use. Thus DLA goes well beyond standard 'hit' analyses (for example, page views and full-text downloads) providing, for instance, data on page view and session time, number of views in a session (site penetration), number of different journals viewed.

More accurate monitoring of information seeking can be obtained because DLA can overcome the worst problems accounted by log analysis outlined below, especially in regard to double counting, accounting for proxy servers and making more effective online time estimates.

It is possible to go beyond usage and furnish detailed assessments of the information seeking behaviour of users, allowing for the profiling of individual user groups. Proprietary software relies on IP numbers to supply user data by country of access and organisational type. DLA provides a wider range of user portraits by: a) identifying, through desk research, which server sub-network labels can be relied upon to yield accurate data on the subject and academic credentials of the user; and b) categorising users by the subject of the journal they viewed.

There are problems associated with log analysis that largely arise from; a) the difficulty of ascribing use to individuals or groups of users, which is compounded by the sheer diversity and size of the user community; and b) the imperfect nature of the log record.

Web logs provide a user 'trace', but not real user or individual identification. Typically all there is to work on is the internet protocol (IP) number, which provides the name of the institution and country to which the user belongs. Logs provide only a partial picture of user activity. A user may well view a number of sites in collecting data to meet an information need. With logs you are usually looking at just one site option. Hence we do not have a record of the user's complete terminal session and views to all sites visited and are viewing just usage on one site. OhioLINK combines journal resources from a number of publishers and aims to provide a one stop shop. However users may in addition visit other web address to collect subject, journal and article information.

Another problem is that robots account for a good deal of usage. Thousands of robots or agents harvest information on the world wide web for a wide variety of purposes – indexing, caching and data mining, for instance. Robots inflate usage statistics by as much as 50 per cent.

Use counts are not completely accurate because of caching, proxy connections, and session definition. Caching has an impact on page views as some viewed pages are not recorded or attributed to a user's search session, leading to more sessions being classified as viewing fewer pages. Caching is the storing of previously viewed pages on to the client's computer; repeat in-session accesses to these pages are made from the cache and are not requested from the web site's server and hence not recorded in the logs, something that underestimates use.

A proxy connection is one where a number of computers are connected to the internet via a single IP number. In such cases, session details of the connected computers are muddled together and it appears that all use comes from the same 'proxy' user since users are identified by IP numbers. This leads to an underestimation of number of users and sessions.

The way sessions are defined can lead to variations in counting. Sessions are sometimes identified in the logs by an identification number. In such cases logs include a session-beginning tag and a session-ending tag, which enables us to make time calculations as well. Unfortunately as far as the logs are concerned, nobody logs off on the web: they just depart anonymously. Typically, then, to estimate a log off – or a session end, and so define a session – a time lapse of inactivity has to be assumed and the industry standard tends to be 30 minutes.

Page view time is estimated by calculating the difference in time between one page and the next page viewed. No estimate can be generated for the last page viewed in a session because there is no log off recorded in the logs.

In addition, double counting can be a problem. If someone views a full-text document in HTML format and then goes on to view this item in PDF, proprietary software tends to count this as two views. This particularly arises where the user comes in from a gateway or third-party site and that site only indexes the HTML version. Hence, when the user clicks through to the article the user is served up with the HTML version. To view the PDF version the user has to come out of the HTML full text version and load up the PDF format version. This process results in the downloading of two items rather than one.

Methods: more on individual library usage data collection

Most libraries do not have electronic journals that are loaded on local servers, so they do not have access to the kind of extensive log data that OhioLink can provide. Instead, they rely on vendor-supplied data and/or data that are collected from other systems in the individual library.

The University of Tennessee (UT) Libraries team has been studying vendor-supplied usage data for its electronic journals. These data are primarily from vendors who supply COUNTER-compliant usage reports. The COUNTER Journal Report 1 (JR1), 'Number of Successful Full-text Article Requests by Month and Journal', provides a summary of usage for individual journal titles for each month and year-to-date. The data points are simply the number of full-text requests made for a particular title in a given time period. There is no information about individual users, where or how they accessed the journals, or what they did with an article after downloading it.

While these vendor data do not provide a profile of the users themselves, they do provide valuable information about the use of library resources. This information allows librarians to ascertain how often the library's electronic resources are being used; what resources are heavily used, and, conversely, not being used; the degree of overlap in journal coverage from multiple sources; and patterns of use over the academic calendar, among other things.

Initial efforts of the team focused on downloading or acquiring UT's 2005 usage data from 29 vendors (some vendors provide reports for more than one product), cleaning up the data to make the data formats consistent, and combining all of the data into one Microsoft Excel file for further analysis. A subset of data for September, October and November 2005 was extracted into a separate file in order to look at UT's usage for the same time period that the readership surveys covered. These data acquisition and preparation activities involved a significant amount of time and effort. In addition to the time spent physically carrying out specific tasks, much time was spent planning the activities and determining the most efficient way to download, manipulate, and compile the data. Combining the data from vendors was a particular challenge because the formats for some data elements were not consistent across all vendors (for example, ISSN with and without a hyphen), and special manipulations of the data were required to be able to sort and compare the data.

A number of basic analyses were done on the data in the Excel file. A simple sort procedure, by total use, provided a look at which journals in which packages were most used during that period. Combining all use of a title (from multiple packages, if more than one) and sorting by total use produced the distribution of use among the individual journal titles, regardless of the source, and highlighted the most used, little used, or unused journals. This facilitated an '80/20' usage analysis, in other words, what percentage of the unique titles accounted for 80 per cent of the usage during that time period.

An important collection management assessment for librarians is looking at the availability of journal titles from multiple sources. For instance, a library may have access to Journal X through a subscription with the publisher, through one or more journal aggregators, and/or from some other type of database or collection. The time coverage may be identical or may differ due to publisher embargoes in aggregator databases, backfile versus current coverage collections, or simply different coverage dates in each package. Sorting the Excel file by title/package highlights the overlap of coverage. Used in conjunction with other subscription information, this may assist the librarian in making decisions to cancel or scale-back on a subscription to reduce redundancy while still maintaining appropriate coverage.

The pattern of usage over time may be plotted for all journals together, individual journal titles, and/or selected groups of titles. From our experience, these patterns generally follow the academic cycle with more use through the mid and late fall and spring semesters (October, November, March, and April), and lulls during breaks, the summer, and early in the semesters. Note that these usage patterns are available only at the monthly level. Discerning patterns of use by day of week, or hour of the day, is not possible with these vendor-reported data because the COUNTER JR1 reports do not include the details of each user transaction.

The vendor data noted above may be combined with other types of data to further characterise the usage of electronic journals. For instance, assigning subject headings to the individual journal titles allows one to see which titles are most used (and, presumably, most important) in different subject areas, or if certain subject areas are more heavily used than others. Similar analyses might involve examining the usage of scholarly and non-scholarly titles. Proportional use by subject also might be an interesting indicator of interest and use: is the subject distribution of the *titles* in a collection consistent with the subject distribution of the *use* of the titles in that collection?

Assigning subject headings to individual titles is not straightforward because many titles may appropriately fall into more than one category and there may be several levels of subject specificity to consider. At the UT libraries we have used the subject headings from our link resolver system to assign the following general subject headings to electronic journal titles: Arts and Humanities/General/News, Business and Social Sciences, Engineering and Technology, Science, and Medicine. In future analyses, we will separate General/News into its own category and create a Multidisciplinary category for titles that do not fit well into just one of the categories above.

Another source of data that supplements the vendor data is our link resolver (SFX from Ex Libris, called FindText at the UT Libraries). Many of the same analyses mentioned above for the vendor data may be done with the SFX data. The vendor and SFX data are not measuring exactly the same thing, but they are likely to reflect the same general patterns of use with regard to high use titles, overlap, and use over time. One distinction of the link resolver system is the ability to record the use of open access journals, which is not available from the vendors' reports. Another advantage is being able, in some cases, to distinguish between the use of older and newer issues of a journal if a journal backfile package is available.

Two statistical reports generated by SFX are of particular interest. The first report, which we generate on a monthly basis to correspond with the COUNTER JR1 reports, provides data elements for how many times a user has clicked on a FindText button for a journal, and which service they chose from the resulting FindText menu. In addition to

providing a measure of how often journal titles are used, the report provides information about specific electronic full-text that users looked for but did not find. For example, when a user chooses a catalogue lookup from the FindText menu, it would indicate that the desired article was not available electronically and the user checked to see if the journal was available in the library collection. Selection of the interlibrary loan service would indicate that the library had no access to the article at all and the user submitted an interlibrary loan request for it. Because the format of this SFX report is very different to the format of the COUNTER JR1 report, we are developing a programming solution for converting the SFX reports into a format that can easily be combined with the vendor data for comparison. Another useful SFX report provides data on how users are getting to electronic journals and articles, for example, from databases, our A-Z e-journals list, or the library catalogue.

Other analyses using vendor-reported and locally collected data and information include the use of subscribed and unsubscribed titles in 'big deal' packages and the use of titles in collections shared with other institutions. In both cases an evaluation may help with collection management decisions. For 'big deals', unsubscribed titles that have high online use may indicate a need for the library to purchase its own subscription or replace a low-use subscribed title of comparable value. Similarly, in shared collections, if one institution has the subscription to a title but does not use it much, and most of the use is coming from one or more of the other institutions, perhaps there should be some subscription adjustments to better align with the demonstrated use of those titles at the different institutions.

Managing the sheer quantity of data generated by vendors and local systems to facilitate analysis and decision-making for electronic resources is a significant challenge. Additional improvements in downloading, reformatting, manipulating, and combining the data are needed to allow desired analyses to become a routine part of the assessment of electronic journal usage. The information we gain about how our investments in electronic journals are benefiting our users, and how we can maximise the use of the collection, may just be worth the effort.

Methods: more on surveys

Surveys can provide demographic information about individual respondents, as well as specific information about the patterns, value, and outcomes of reading and how those may vary with characteristics of users. They are particularly helpful in focusing on actual readings, rather than merely on downloads, and for finding out how useful a particular reading was for its purpose. Surveys can be used to get a picture of journal article reading from all sources, including both print and electronic and both from the library collection and other sources, including personal subscriptions. The value of survey results is, of course, reliant on response rate and we assume that the respondents are representative of the user group as a whole. Although, unlike log data, they rely on self-reported data, and careful consideration of the types of questions asked can help improve the reliability of results.

In separate surveys, both faculty and students of five US universities and two Australian universities were asked three types of questions:

1. Demographic questions about themselves, including age, gender, status (year in school or rank). In addition, faculty were asked about their publication record, number of personal subscriptions, and whether they had received any awards in the past two years. Demographic questions specific to this study include percent of effort on various work tasks (such as teaching, research, administration, etc.), number of publications in the last two years, co-authorship and funding for the last article published, awards or special recognition in the last two years, and number of personal subscriptions. Each of the reader-related questions can be used as an independent variable and the population is the total number of academic staff at the university.
2. Respondents were first asked to recollect how many scholarly articles they read in the last four weeks. Scholarly articles are defined as 'those found in journal issues, web sites, or separate copies such as preprints, reprints, and other electronic or paper copies'. Reading is defined as 'going beyond the table of contents, title, and abstract to the body of the article'. Respondents were asked to remember just this short period to improve the chance of accurate recollections. General recollection questions were held to a minimum, as they are often difficult for a respondent to answer them accurately.
3. Details on the last article read. These questions focused in detail on the specific article most recently read, a technique that improves the accuracy of the respondent's ability to remember and allows complex analysis of the value and purpose of reading. Questions about how the article was discovered, its format, time spent reading, the purpose of reading, and the value of the reading to the purpose, provide information about the relative role and value of library collections or alternatives in the overall reading patterns of academics. The incident of last reading is a variation of the critical incident technique, where the universe is all readings by academic staff within the last month. A two-stage sample is taken: the first stage is the readers and their total readings in a month, and the second stage is one incident of their most recent reading, which is assumed to be random in time. Analysis of the reading-related questions (incidence of last reading) allow conclusions about readings rather than readers, answering such questions as, for example, does the value of the reading or the average time spent per reading vary with the purpose of the reading?

Preliminary findings

Findings: deep log analysis

Based on the study of on-campus use at four Ohio universities over 15 months the following analyses were undertaken in order to profile and compare information seeking behaviour:

1. **Number of pages viewed, over day of the week and month of the year.** Altogether 2,250,000 pages were viewed and 339,000 sessions conducted, demonstrating how genuinely popular e-journal databases are with the scholarly community. There were huge differences between universities and this would appear to be largely a function of research activity and the size of the academic community – the more research active and larger the university the higher the activity level. Monthly use is highly variable and linked to the rhythms of the academic year, with peaks being achieved in the autumn and spring terms and lows recorded in the summer and Christmas breaks. The monthly patterns differed between the research universities and the teaching universities with the latter showing less pronounced peaks and troughs. There were big differences too in usage by day of the week – Tuesday recorded the highest usage (19 per cent) while Saturdays recorded the lowest usage (7 per cent).

2. **Number of sessions conducted, per day.** The number of daily sessions varied quite considerably from about 200 to 1000. Again there was a falling off in sessions during the university holiday breaks especially in the summer and winter breaks. There was no discernable increase in the number of sessions recorded over the 15-month survey period. In addition to the rises and falls of the academic year there were the rises and falls in use during the week, peaks during weekdays and troughs at the weekends. The research universities accounted for a mammoth (95 per cent) of them, with the most research extensive university accounting for 60 per cent itself.
3. **Number of pages viewed in a session (site penetration).** About one in five users made one view in a session, a little under a third made two to three views, over a third made four to ten views, a tenth made 11 to 20 views, while just 6 per cent made more than 20. In the case of this 'activity' metric, interestingly, it was the smallest teaching university that had the most active sessions, not the most research extensive university. Thus 60 per cent of the former's sessions saw four or more pages viewed, whereas the equivalent figure for the research extensive university was 48 per cent. Indeed, over a quarter of sessions in the small teaching university saw more than ten pages viewed in a session.
4. **Time spent viewing online.** It was found that 32 per cent of sessions lasted under three minutes, 15 per cent lasted between three and seven minutes, 15 per cent lasted between seven and fifteen minutes, and over 38 per cent recorded session times of greater than fifteen minutes. There were large differences between the four universities – as much as 300 per cent – but the difference was not accounted for by their research activity. The time taken (median) to view a page was around 20 to 22 seconds and this was the same for all the universities. There was a significant increase of about 20 per cent in page view time over the 15-month survey period. Article view time was significantly longer at the research universities (89 seconds) compared to the two teaching universities (76 and 68 seconds).
5. **Type of page viewed (article, abstract, TOC etc).** The pages most frequently viewed were lists – lists of journals, articles, subjects and so on, something which demonstrates the amount of navigating to content that goes on in very large and complex databases. Not far behind were views to the full-text of the article. There were some significant differences between universities, with the teaching universities recording relatively more views to abstracts and search pages.
6. **Top journals viewed.** Over the fifteen month period 13 per cent of journal titles recorded just one view, 58 per cent of journals were viewed from 2 to 20 times, 23 percent of titles were viewed 21 to 100 times, just 5 per cent of journals were viewed 101 to 500 times and under half a per cent of them were viewed over 500 times. Of the 6,000 or so journals on offer just 20 accounted for between 11 and 19 per cent of use, with the largest university, perhaps unsurprisingly, showing the least concentration in use. Given the subject and research interests of the four universities it was not surprising that the top ten titles for each university differed significantly. No journal was represented in all the lists; one journal (*American Psychologist*) was represented in the lists of three, and six journals were in the lists of two universities.
7. **Publication year of pages viewed.** There was a focus of interest in the most recent articles. Twenty-two percent of pages viewed were published in the current year, 42 per cent were one to three years old, 24 per cent were four to seven years old and 12 per cent were over seven years old. In terms of the four universities, distributions were fairly similar, with the exception of the small teaching university whose liberal arts roots showed up in its interest in older journal articles – 49 per cent were four years or older, as compared to a figure of 37 per cent for the research extensive university which had strengths in science.
8. **Navigational approach adopted (lists/menus, search engines).** Those people using the internal search engine were more likely to view an abstract as well as an article in their session. The explanation for this difference lies in the fact that search engine users have a greater number of items to view and they resort to abstracts to make a quick selection. It was also the case that users employing the search engine were far more likely to view a wider range of journals.
9. **Number of journals viewed in a session.** With so many journals on offer one might have expected users to avail themselves of quite a few titles. In fact, in terms of the proportion of sessions where a user viewed two or more journals within the four universities this varied between 60 per cent (medium size teaching university) and 42 per cent (research extensive university).
10. **Subject of journals viewed.** There were big differences between subjects. Thus health and medicine journals recorded the highest page views, and just fewer than 500,000 views were made to this subject. About 300,000 chemistry related pages were viewed and 195,000 social science related pages were accessed. Computer science, earth science and mathematics pages were the least viewed. In terms of the distribution of the number of page views in a session social science and physics subjects viewed a greater number of pages in a session. About two thirds (66 per cent and 71 per cent) viewed four or more pages in a session. Users viewing health and medicine were least likely to view four or more pages: just 49 per cent did so. In terms of session time, those viewing social science journals recorded the longest sessions – 71 per cent recorded sessions lasting three or more minutes and 53 per cent recorded sessions lasting over fifteen minutes. Arts and humanities scholars conducted the shortest sessions with only 63 per cent of these sessions lasting more than three minutes.
11. **Staff/student use.** In the case of one university it proved possible to identify with a reasonable degree of accuracy staff and student sub-networks. The analysis confirmed the long-held belief that students were the majority users of digital resources, there are after all more of them. Student use was about ten times that of staff. However, in terms of the number of views in a session, the staff made a greater number views in a session compared to the student labelled network: 27 per cent of sessions saw eleven or more pages viewed in a session; however, just 12 per cent of student sessions viewed that many pages. Other differences between staff and students were: a) staff accounted for a much high proportion of social science use; b) in terms of day of week staff made up a lower percentage of use at weekends; c) staff were more likely to use the search facility; and d) staff were less likely to conduct sessions where just current and recent journals were viewed.

Findings: individual library usage data collection

There are a variety of sources of usage data for individual libraries, from vendor reports to data gathered within the library, and the data tells us about different aspects of our e-journal use. Some findings from UT's vendor and link

resolver reports provide insights that are likely to be applicable to other libraries.

1. **Impact of the academic year.** Specific courses and course scheduling can have an appreciable impact on usage statistics. For example, *Billboard* and *Rolling Stone* are among the most used electronic resources in the library collection, most likely as a result of the popular 'History of Rock' class. Librarian demonstrations and instruction can also have an impact on usage statistics. The primary resource in a series of instruction classes will tend to have higher levels of use during the term.
2. **Subjects of journals used**. Vendor data showed that journals in the sciences accounted for the largest percentage of use (27 per cent). Medicine and health accounted for 19 per cent of the use, and engineering and technology for 10 per cent, meaning that all of the science and technology subjects combined accounted for 56 per cent of electronic journal usage. The social sciences received 23 per cent of the use, and arts and humanities/general received 21 per cent.
3. **Scholarly versus non-scholarly use.** A number of the UT libraries highest use titles are considered non-scholarly (for example, *USA Today* and *Rolling Stone*). For these titles, it is difficult to tell if they are being read for personal or scholarly use. As noted previously, *Rolling Stone* is regularly used for a class assignment, but is surely also read for personal enjoyment. *USA Today*, *Newsweek*, and the *New York Times* are used for class papers and speeches, particularly by undergraduates, but are also read for current awareness.
4. **Journal packages make a difference.** An analysis of the use of all electronic journal titles from September to November of 2005 revealed that 11.5 per cent of the titles accounted for 80 per cent of the use. Much of this can be accounted for by the large number of zero-use titles.
5. **A close look at embargoed titles can help our users.** A fair number of electronic journals in aggregator packages are embargoed, with, for example, the most recent year being unavailable. By combining SFX request data with vendor use data for such titles, we can discover when users are requesting recent embargoed articles from interlibrary loan, and can consider the need for a subscription with more complete coverage.
6. **Backfiles can be evaluated with SFX data.** Usage data supplied by vendors do not distinguish between use of current issues and back issues that are part of a backfile package. The SFX data, on the other hand, allow us to make that distinction when the backfiles are listed as separate targets.

Findings: surveys

Surveys reveal information on both overall patterns of use and what factors make a difference in patterns of use. The surveys reveal a variety of user-related insights, including:

1. **Amount of reading and time spent reading.** Faculty report reading, on average, 23 articles per month (standard deviation = 31.273), while students report reading on average 15 per month (SD = 20.217). Faculty spend an average of 33 minutes per reading (SD = 30.56) and students 36 minutes per reading (SD = 35.598). Although the average time spent per reading is short, the total amount of time devoted just to reading journal articles (and not including searching for and locating the articles) represents a considerable amount of time every month. These numbers tell us about individuals and reading from all sources and both print and electronic. To compare them with log data we would need to take the subset of only those readings that are from the library's electronic collection and then extrapolate up to the entire student body population.
2. **Source and format of reading.** Faculty and students find and read articles from the library, from personal subscriptions, from colleagues, by following citations in other articles, and from other sources. Nearly two-thirds of these readings are from an electronic source, although most are still printed out for easier reading. Students read articles on the screen more often than do faculty members do (18 per cent of faculty readings are onscreen while 30 per cent of student readings are onscreen).
3. **Purpose and value of reading.** Articles are read by faculty for many reasons including, in order, research, teaching, current awareness and writing. Students read most often for course assignments and thesis or dissertation. On the whole, articles are rated most often as somewhat important or essential to the purpose. When asked to comment specifically on the value of the last article they read, faculty report that, most often, readings inspired new thinking, improved results, or changed, broadened or narrowed the focus of their work.
4. **Subject discipline makes a difference.** The subject discipline of faculty respondents is the most consistent predictor of differences in reading patterns. Medical faculty, for example, read significantly more journal articles on average (35 per month as compared to 13 per month on average by humanities faculty) and spend less time per reading (25 minutes per reading on average compared to 37 minutes per reading for humanities faculty or 43 minutes on average for engineering faculty.) Medical faculty read more often from personal subscriptions and print sources.
5. **Age and status may make a difference.** Among students, their status – whether they are undergraduate or graduate/post-graduate – makes the most difference in how much they read and why they read. Not surprisingly, graduate students read more articles on average and read more for research (thesis or dissertation) than do undergraduates. Since student status is related to age, it is not surprising that age also makes a difference. For faculty, age makes a difference in some reading patterns. Faculty members over 40 are significantly more likely to get information from a print source than the younger faculty. Faculty readers aged 31 to 40 are significantly more likely to read for research than the other age groups. Reading for research declines significantly after the age of 50, when reading is more likely to be for teaching. Younger faculty readers read significantly fewer articles in the first year of publication than do older readers.

Conclusion

Libraries and publishers have many ways to gather information about the use of their collections. Some of these data can also tell us about users and patterns of usage. No single method of gathering usage data is perfect, and each provides unique views of users. Deep log analysis tells us what large numbers of users across multiple institutions view and download. Deep log analysis can reveal patterns of use and can be used to compare usage on a large scale. Vendor and collected data reveal patterns of use specific to a library and its collection. They can be used to monitor changes in the collection or curriculum of a specific institution and show how users react to the decisions and rhythms within their university. Finally, surveys can be used

to discover readings beyond electronic library collections. Surveys can be used to put use of a library's e-collection in perspective with the total readings, print and electronic, of university faculty members and students. Unlike any type of log analysis, surveys can get at motivation, purpose, and explicit values of use, although they rely on self-reporting and a sufficient number of respondents.

The level of effort (and therefore cost to the library) required by each of these methods varies as well and is not insignificant. Even COUNTER-compliant vendor reports require some effort in processing and comparing across systems. Together, these methods provide many insights into how our collections are being used and, over time, they reveal how usage patterns are changing. Ultimately the goal is to build collections and services that meet the needs of users. Understanding those needs and usage behaviours is an essential step in making that happen.

Acknowledgements

This project was funded by the US Institute for Museum and Library Services. The project team also includes Donald W King and graduate students Lei Wu, Liuyan Wang, Karen McClanahan and William Birch.

References

Baker, Gayle, and Eleanor J. Read. 2008 [in press]. Vendor-supplied usage data for electronic resources: A survey of academic libraries. *Learned Publishing.*

Baker, Gayle, and Carol Tenopir. 2006. Managing the unmanageable: Systematic downloading of electronic resources by library users. *Journal of Library Administration* (Library/Vendor Relationships) 44 (3/4): 11–24.

Huntington, Paul, David Nicholas, and Hamid R. Jamali. 2007 [in press]. Web site usage metrics. *Information Processing & Management* 43.

Huntington, Paul, David Nicholas, Hamid R. Jamali, and Ian Rowlands. 2006. Article decay in the digital environment: A usage analysis by date of publication employing deep log methods. *Journal of the American Society for Information Science & Technology* 57 (13): 1840–51.

Huntington, Paul, David Nicholas, Hamid R. Jamali, and Anthony Watkinson. 2006. Obtaining subject data from log files using deep log analysis: Case study OhioLINK. *Journal of Information Science* 32 (4): 299–308.

Manoff, Maribeth, Eleanor Read, and Gayle Baker. 2006. All that data: Finding useful and practical ways to combine electronic resource usage data from multiple sources. In *Building Effective, Sustainable, Practical Assessment: Proceedings of the Library Assessment Conference*, held September 25–7, 2006. In Charlottesville, Virginia, edited by Francine DeFranco et al., 409–15. Washington, D.C.: Association of Research Libraries, 2007.

Nicholas, David, and Paul Huntington. 2006, June. Digital journals: Are they really used? *Interlending and Document Supply* 34 (2): 74–7.

Nicholas, David, Paul Huntington, Tom Dobrowolski, Ian Rowlands, Hamid R. Jamali, and Panayiota Polydoratou. 2005, Summer. Revisiting 'obsolescence' and journal article 'decay' through usage data: An analysis of digital journal use by year of publication. *Information Processing and Management* 41 (6): 1441–61.

Nicholas, David, Paul Huntington, and Hamid R. Jamali. 2007 [in press]. Diversity in the information seeking behaviour of the virtual scholar: Institutional comparisons, *Journal of Academic Librarianship* 33.

Nicholas, David, Paul Huntington, and Hamid R. Jamali. 2007 [in press]. User diversity: As demonstrated by deep log analysis. *Electronic Library* 25 (3)

Nicholas, David, Paul Huntington, Hamid R. Jamali, and Carol Tenopir. 2006, March. Finding information in (very large) digital libraries: A deep log approach to determining differences in use according to method of access. *Journal of Academic Librarianship* 32 (2): 119–26.

Nicholas, David, Paul Huntington, Hamid R. Jamali, and Carol Tenopir. 2006, July. OhioLINK – ten years on: What deep log analysis tells us about the impact of big deals. *Journal of Documentation* 62 (4): 482–508.

Nicholas, David, Paul Huntington, Bill Russell, Anthony Watkinson, Hamid R. Jamali, and Carol Tenopir. 2005, October. The big deal: Ten years on. *Learned Information* 18 (4): 251–7.

Tenopir, Carol, Gayle Baker, Eleanor Read, Maribeth Manoff, Kitty McClanahan, David Nicholas, and Donald W. King. 2007 [in press]. MaxData: A project to help librarians maximize E-journal usage data. In *Usage Statistics of E-Serials,* edited by David C Fowler. Binghamton, NY: Haworth Press.

Tenopir, Carol, and Donald W. King. 2000. *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers.* Washington, DC: Special Libraries Association.

Tenopir, Carol, and Donald W. King. 2004. *Communication Patterns of Engineers*. NJ: IEEE/Wiley-InterScience.

Tenopir, Carol, and Donald W. King. 2007 [in press]. Perceptions of value and value beyond perceptions: Measuring the quality and value of journal article readings. *Serials.*

Contact

Carol Tenopir
ctenopir@utk.edu