Journal of Web Librarianship 2009, v. 3, n.1, pp.3-14.
ISSN: (Print 1932-2909) (Online 1932-2917)
DOI: 10.1080/19322900802660292
http://www.taylorandfrancis.com/
http://www.informaworld.com/smpp/title~db=all~content=t792306922~tab=issueslist
http://www.informaworld.com/openurl?genre=article&issn=1932-2909&volume=3&issue=1&spage=3

Web Analytics: A Picture of the Academic Library Web Site User

ELIZABETH L. BLACK

This article describes the usefulness of Web analytics for understanding the users of an academic library Web site. Using a case study, the analysis describes how Web analytics can answer questions about Web site user behavior, including when visitors come, the duration of the visit, how they get there, the technology they use, and the most popular content using an open-source Web log analysis tool. The author offers suggestions for future research into user motivations to complement the findings possible using Web analytics.

INTRODUCTION

Web analytics are used widely by commercial companies and marketing departments to assess the effectiveness of their Web sites (Sterne 2002). Web analytics, known also in the literature as Web metrics, Web log analysis, and Web statistics, are the objective tracking, collection, measurement, reporting, and analysis of quantitative Internet data to optimize Web sites (Kaushik 2007). This information is found in the Web server logs generated by the use of the Web site and interpreted by a software application, such as Web Trends or AWStats. These Web analytic tools create reports using the Web server logs in order to make the information easier for people to use. Many libraries limit their use of this data to reporting items required by the national reporting entities for comparing libraries, such as the Association of Research Libraries and the National Center for Education Statistics. Beyond reporting numbers of visitors to the Web site and the number of searches of electronic resources, Web analytics offers other valuable information describing use of the library Web site. This information is already in the logs of the Web server hosting the library Web site, so it is an inexpensive and readily available source to begin learning more about the library's online patrons.

This article describes the user information found from an analysis of two years of Web site statistics of the Ohio State University Libraries' Web site. It uses a case-study approach to demonstrate the use of Web analytics to learn more about users of academic library Web sites. The author believes the best Web sites and library services are those that keep their users in mind as they are built and maintained. This article shows one source of data from which a library can learn about its Web site users and suggests ways to use that information to improve the site.

This case study will analyze the Web site logs from the OSU Libraries in an attempt to answer the following questions:

- When do visitors come to the Web site? How long are they staying?
- How do users get to the Web site?
- What technology do the Web site's users have?
- What is the most popular content on the Web site?

LITERATURE REVIEW

Use of Web Transaction Logs

David Nicholas has written a number of articles with a variety of authors on the utility and application of transaction log analysis (Nicholas et al. 2003; Nicholas and Huntington 2003; Nicholas et al. 2000; Nicholas et al. 1999; Nicholas et al. 2002). He noted Web usage logs offer a direct and immediate record of what people have done on a Web site—not what they say they might do but what they have actually done. This record makes the logs invaluable for understanding a site's users.

There are difficulties with using Web server logs. The data is the record of one machine's interaction with other machines because server logs are designed to record requests the Web server receives and the manner in which it responds. Therefore, it records basic facts about the interaction, such as the time it occurred and the IP address, the unique address of computers on the Internet from which the request came. The IP address of the requesting machine is sometimes the actual machine of the user and sometimes the IP address of the Internet service provider's proxy server. A proxy server works as an intermediary between the Internet and a smaller network of computers. For studies looking closely at the place of origin of the use, this is a quality issue requiring additional work to mitigate, as Nicholas and Paul Huntington did with a system called micro-mining (Nicholas and Huntington 2003).

Another issue with Web server logs comes from an operating principle of the Internet itself: caching. In order to reduce traffic on the Internet, many Web access providers and personal computers cache files; that is, they store the page in a temporary folder, and if a Web page is revisited within a predetermined timeframe, the local or cached version of the page is given to the user instead of the one on the Web server hosting the site. This makes the delivery of the Web page more efficient but does make the usage statistics kept on the Web server less accurate.

Sanghamitra Jana and Supratim Chatterjee (2004) used cybermetrics to analyze Web site's content, applying bibliometric principles to the data found in the Web server logs. They evaluated hits, page views, and visits, but found hits are not very informative because they can vary widely because of site's graphical design and architecture instead of the actual usage. They found the best measure was user sessions, which are labeled as visits in some Web analytic software tools.

Learning about User Behavior with Web Log Analysis

Looking into server logs to learn about user behavior with systems is not new. Libraries were early adopters of this technique with catalog systems, as noted by T. A. Peters in a *Library Hi Tech* issue devoted to the topic (Peters 1993). As the library Web site becomes a key service delivery point, understanding user behavior on the site increases in importance. An analysis similar to the one shared in this article becomes an essential part of the library evaluation tool kit; as Oliver Pesch noted, "Statistics are a measurement of users' actions that we try to correlate to their intentions" (2004). Other means must be used to determine intentions, and statistics are a starting point to begin asking the questions.

Several articles have reported on using Web usage statistics to assist in evaluating services delivered via the Web. The ARL E-Metrics project explored ways to add outcome

measures to library assessment; one outcome suggested was measuring Web site usage (Fraser and McClure 2002). Jeanie Welch (2005) wrote on the need to incorporate Web usage statistics into the data more traditionally kept by libraries. She concluded that compiling and analyzing server statistics of library Web sites has several advantages for librarians, including demonstrating the effectiveness of reaching patrons remotely. A. D. Phippen (2004) proposed using web analytics to evaluate the behavior and usage of virtual communities in the social sciences.

Another reason to analyze Web site usage statistics is to assess the effectiveness of the site itself. Theresa Murdock (2002) described the work at the University of Washington Libraries to create a more user-centered ready reference Web page through an examination of both server statistics and e-mail reference queries. An Online Computer Library Center (OCLC) report on perceptions of college students found most college students know the library Web site exists, but many do not use it because they feel other Web sites have better information (DeRosa et al. 2006). It is critical that academic libraries use all the tools available to them, including Web analytics, to make the library Web site a valuable resource to college students.

An effective model for analyzing Web site statistics was put forth by Laura B. Cohen (2003a, 2003b). She outlined the pros and cons of using Web site log statistics for analysis, adding a caution that statistics for dynamically generated pages should not be counted the same way as requests for static pages. The most valuable part of Cohen's work for this study was the outline of which data elements to collect and their suggested analysis. Cohen noted for each category of data a rationale for collecting that group, including what valuable information can be interpreted from the grouping. The analysis of the Ohio State University Libraries' Web site statistics aligns with Cohen's model focusing on static Web pages. It describes one possible method of using Web analytics with a focus on learning more about the users, and is unusual in that it applies the tools described by the other works in a concrete manner, making it easier for others to repeat.

METHODOLOGY

AWStats is a log file analysis program that offers a graphical display to the data in all major Web server log files and is a freely available open source application. Full documentation is available at http://awstats.sourceforge.net/. The host of the Ohio State University Libraries' Web site, the Ohio State University Office of Information Technology, offered AWStats as the Web analytics tool. Because of its availability and the Libraries' interest in using open-source tools, this study used AWStats. Several other Web analytics packages, such as Web Trends and Google Analytics, could easily be used to replicate this study.

This study analyzes Web usage statistics for the Ohio State University Libraries' Web site for a two-year period, from January 2005 through December 2006. The study starts with January 2005 because that is the first month the entire Web site was available at the current address: http://library.osu.edu. The period of 24 months was selected to give enough data and time to identify trends in the site usage.

All data came from the reports generated by AWStats. While there are those, such as David Nicholas, who recommend going straight to the server logs for the most control over the analysis (Nicholas et al. 2000), this author relied on AWStats to pull together the data from the logs. This makes the study more easily replicated by librarians at other institutions. The tools are also getting more accurate in bringing the data from the logs (Kaushik 2007).

Despite the increasing reliability of today's Web analytic tools, the cautions noted earlier about Web statistics must be remembered. The purpose of the logs is to record communications between two machines: the computer of the user accessing the Web site and the Web server on which the site resides. As such, the motivations of the users are unknown.

Data from the AWStats reports were entered into an Excel spreadsheet and graphed and sorted to identify patterns of the use over time. The most meaningful number available was used to answer each question; for overall site usage, "visit" is the term used. One "visit" equals all of the pages viewed by one unique IP within an hour; this is sometimes known as a session. For tracking the use of technology, such as browser or operating system, "hits" is used. Hits are not meaningful for overall use because a hit reports a request for a file to the Web server and can misrepresent content usage because most content pages contain more than one file, but for tracking the technology use of sites' users, "hits" adequately serves the purpose.

ANALYSIS

Overall Usage

The Ohio State University Libraries' Web site is heavily used. Over the survey period, the site had an average of 304,202 visits per month. One visit is the viewing of any and all Web pages by a unique IP address. Usage overall increased. Total visits in January 2005 were 229,336, and the total visits in December 2006 were 347,517.

The number of repeat visits declined over the survey period. A repeat visit is the total number of visits minus the number of unique visitors. A unique visitor is a host, determined by IP, which requests at least one file from the Web server for a given month. The average number of visits per visitor in January 2005 was 2.17, the highest average for the entire period. The average number of visits in December 2006 was 1.75. The lowest average for the entire period was 1.61 in December 2005. In 2005, six of the months had an average visits figure of 2.0 or above; in 2006, only two months had an average above 2.0. This shows users are not returning to the Libraries' Web site as often as they had in the past. Statistics cannot tell why this is so; further research is needed to answer that question. But this could represent a troubling trend, indicating that perhaps the Web site is not meeting the information needs of its users, causing them to not return.

Timing and Duration of Visits

The site is visited most heavily on Mondays, Tuesdays, and Wednesdays. Monday was the day most often in the top spot for visits. The hours of heaviest use were 2 p.m. and 3 p.m. ET, showing that patrons use the site during traditional work hours. Since the vast majority of visitors to the site came from IP addresses registered in the United States, it is safe to assume these visits were during a workday in a U.S. time zone. This data informs those who maintain the computers running the Web site to avoid system changes or other work during these peak hours and days.

Most visits to the site have a very short duration: less than 30 seconds (see Figure 1). Each month of the survey period, more than 60% of the visits to the Ohio State University Libraries' Web site were for 30 seconds or less. In all but four months in 2005, the duration of 80% of the visits was five minutes or less. This supports reports by others that users move very

quickly from one Web page to another, often just clicking on the first thing that draws that user's attention (Krug 2006). A user in one study put it sufficiently when he said, "I don't think, I click" (Novotny 2004, 530).

The data reports that users are leaving the OSU Libraries' Web site quickly, but it does not describe their motivation: if they left because they found what they wanted—such as the link to the library catalog prominently featured on the home page—or because they did not find what they needed in a quick review of the page on which they landed. The content and design of a Web site used in this way must be tested with real users in order to be truly customer centered. The next step is to conduct usability tests with users to see if the presentation of the material on the site meets their needs.

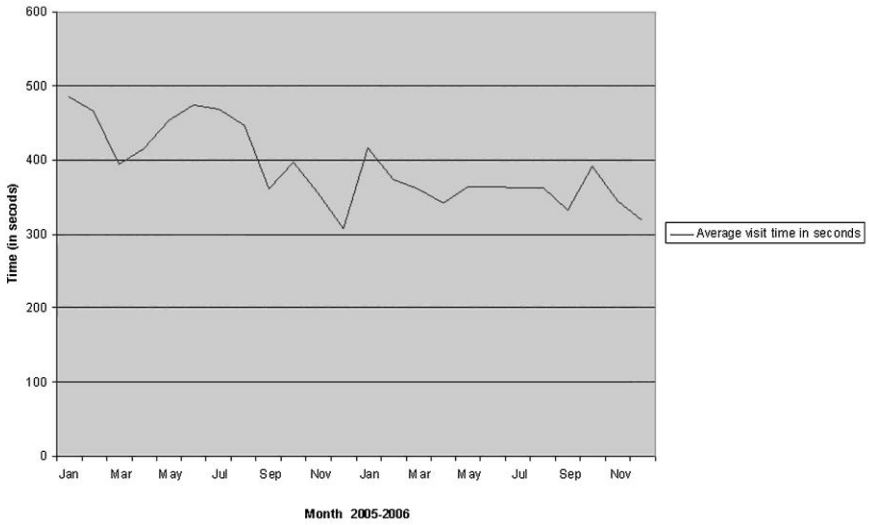

FIGURE 1 Average visit duration.

How Users Arrive at the Web Site

By far, most users access the Ohio State University Libraries' Web site by typing in the address or following a bookmark (see Figure 2). The percentage of users who access the site this way increased dramatically over the survey period, from 64.2% in January 2005 to 86.8% in December 2006. January 2006 was the month with the highest direct access, at 90.3%. The percentage of direct access remained in the 80% group for the entire second half of the survey period. At the same time, access via search engines dropped from a high of 20.5% in March 2005 to 5% or less for the majority of 2006. Connections to the Web site from external page links, the final method of connection available, also decreased during the survey period. It was highest in January 2005 at 15.9% of connections and 5% at the lowest point in December 2006. The external sites consistently delivering users to the Libraries' site are the University Web site and the Library catalog server.

Decreasing access to the OSU Libraries Web site through search engines and external links is of great concern because, as OCLC research shows, most college students describe search engines as a better match for their lifestyle (De Rosa et al. 2006). Students regularly begin information searches with a search engine rather than the library Web site; they do not go directly to known sites until they feel they need to do so. The future success of the site will

depend on being where the users are—the search engines.

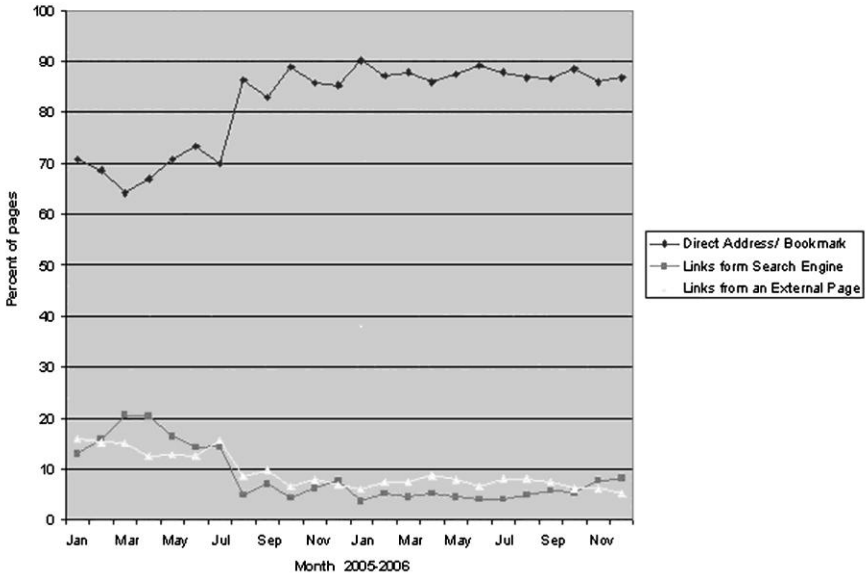

FIGURE 2 How users arrive at the Web site.

One recommendation from this analysis is to make the Web site friendlier to search engines. Lorcan Dempsey (2007) made a strong argument that the trend is to create content that is easily aggregated so it can be added to a user's workflow. This is essential, Dempsey noted, because users have scarce attention to spare; they need the library to make the valuable content easy to grab where and when it is needed. The good news is that the percent of access by search engines increased slightly at the end of the survey period, to 7.6% in November 2006 and 8.15% in December 2006.

Technology of Users

The Web browser used by most visitors to the Ohio State University Libraries' Web site is Internet Explorer (IE); however, the use of IE decreased over the survey period (see Figure 3). In January 2005, IE was used for 91.8% of the hits on the OSU Libraries' Web site. By the end of the survey period, IE sent just 73.6% of the requests. This parallels similar usage patterns of IE on the Web in general, as reported by Net Applications (2007). While the use of the Firefox browser did increase during the survey period, it did not increase enough to account for all of IE's loss of use. The greatest increase in the reports from AWStats was in "unknown browser." This leaves the analysis incomplete—while it is clear IE use is decreasing, it is not obvious to which browser users are moving. The 73% use of IE is still significant; therefore, all parts of the Web site must be thoroughly tested with IE before they are made available to users.

Windows is the operating system most widely used by visitors to the OSU Libraries' Web site. Web analytics reports show a range from a high of 96% of usage in January 2005 to the lowest reported usage of Windows, 78%, in September 2006. The Macintosh operating system is the second most widely used, ranging in usage from 2.5% to 3% throughout the survey period. Since the majority of visitors use Windows, the Libraries must make sure all things offered via the Web site are compatible with Windows.

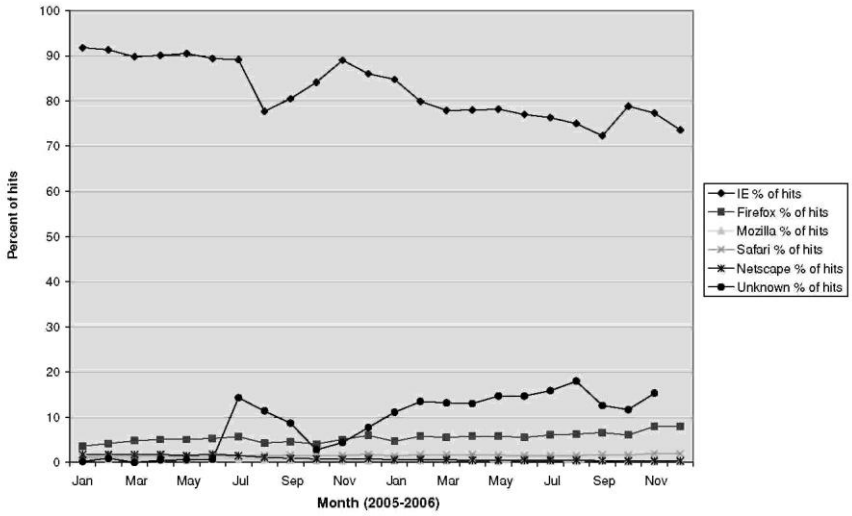

FIGURE 3 Browser usage.

Reviewing the version of browser and operating system usage demonstrates the rate at which users adopt new technologies. The release of Internet Explorer 7 on October 18, 2006, was shown in the Web usage statistics—although it was a small amount, with only 1% of IE requests coming from version 7. By December 2006, the last month of this analysis period, 9% of IE users were using version 7. This is lower than the 18.26% of use of IE7 by the general Web population, as reported by Net Applications Market Share (Net Applications 2007). This suggests that users of the OSU Libraries' Web site are slower than general Web users to adopt new technologies. Heavy use of campus-wide software applications, such as the course management system, and of computers supplied and supported by the University may be the reason for the slower adoption of newer browser versions than the general population.

Most Popular Content

The Web page with the most views is consistently the home page of the OSU Libraries' Web site; it is also consistently the top entry page for Web site visitors. This aligns with the data that shows most users arrive at the site either by typing the address directly into their browser or following a bookmark.

The citation guides created by the Libraries' instruction office are also consistently among the top ten most viewed pages. The Chicago, MLA, APA, and Turabian Style citation guides were in the top ten most months of the survey, with the Chicago and APA guides in the top ten each month. These pages are also high on the entry and exit pages lists, demonstrating that many users come to the Web site specifically to view this information.

An analysis of the words and phrases used on search engines that in turn lead users to OSU Libraries' Web pages showed the citation guides are frequently requested material. The majority of the terms and phrases that totaled 1% or higher of the total search terms each month related to citation styles and manuals. The second category related directly to the Ohio State

University or the Libraries.

Other Web pages in the top ten most visited pages are the list of libraries, library hours, and the Web page describing how to find information in general and links to alternate catalogs. These are more way-finding or navigating Web pages that lead users to other pages that contain the content they seek. Two library locations also regularly made the list of the top ten: the Science and Engineering Library and the Music Library.

During the height of football season each year of the survey period, the pages hosted and maintained by the University Archives about the football game between OSU and Michigan took over the top ten, the top entry, and top exit page lists. In November 2006, four pages related to this event were in the top ten, and five were the top entry and exit pages. Users found this information—most likely through search engines—read it and then left the site. Spikes in overall usage of the Web site can also be traced to traffic generated by the interest in this football rivalry.

A few other Web pages with specific content broke the top ten for one to three months of the survey. These were, in other months, a bit further down the list. Pages with subject information such as the subject gateway to the paid databases and University Archives pages about Jesse Owens and the polar expeditions were popular. It is perhaps significant that content such as this made the entry list more than the top ten viewed list, suggesting compelling, unique information is a draw and will bring visitors to the Libraries' Web site. Further research to test this hypothesis is recommended.

In many cases, the top entry and exit pages were the same as the most visited Web pages. This makes sense for the content pages. The viewer finds the information they seek in the content, such as the heavily used citation guides, and leaves the site. The Libraries' Web page met the information need—worked in the flow of the user, as Dempsey (2007) might say—and the user left, hopefully, happy and satisfied. However, usage statistics alone cannot report if users are satisfied; further research to test this hypothesis is recommended.

CONCLUSIONS

The high number of visits suggests the OSU Libraries' Web site has valuable content; further study is needed to prove if this is true. High visit numbers show many people are viewing the content, thus increasing the likelihood that it is valuable and reaching those to whom it will be useful. Based on this premise, the author suggests this content be made even more accessible to those who do not know it is there. One suggested strategy is the creation of a sitemap optimized for the major search engines. The anticipated result will be increased visits from users via search engines. The second strategy is an analysis of the next group of highly visited pages to determine what content might need to be moved to a more prominent location on the site.

An analysis of the Web log data is not complete without usability studies to supplement the findings with observations of user behavior. The statistics are a trail left by the user, but they do not explain the motivations behind that behavior. A usability study that involves observations of the target audience using the Web site will investigate why users made the choices evident in this analysis. Further, a usability study will focus on the local users, something impossible to do with the Web logs.

Google's entrance into the Web analytics field with their Google Analytics tool invigorated work in this area. The author suggests exploring other methods of gathering usage data, such as using JavaScript tags and combining this quantitative data with qualitative data, like

surveys, to achieve the goal of combining the why—the intent and motivation of the users—with the what—the observable behavior of the users. This complete picture will lead to solid actionable objectives for serving library patrons through the Web site.

REFERENCES

Cohen, Laura B. 2003a. A two-tiered model for analyzing library Website usage statistics, part 1: Web server logs. *portal: Libraries and the Academy* 3(2): 315– 26.

———, 2003b. A two-tiered model for analyzing library Website usage statistics, part 2: Log file analysis. *portal: Libraries and the Academy* 3(3): 517– 26.

Dempsey, Lorcan. In the flow: From discovery to disclosure. Paper presented at CIC Library Conference, University of Minnesota, Minneapolis, MN, March 19, 2007. http://www.oclc.org/research/presentations/dempsey/cic.ppt (accessed May 5, 2008).

DeRosa, Cathy, Joanne Cantrell, Janet Hawk, and Alane Wilson. 2006. College students' perceptions of libraries and information resources. Dublin, OH: OCLC Online Computer Library Center. http://www.oclc.org/reports/perceptionscollege.htm

Fraser, Bruce T., and Charles R. McClure. 2002. Toward a framework for assessing library and institutional outcomes. *portal: Libraries and the Academy* 2(4): 505–28.

Jana, Sanghamitra, and Supratim Chatterjee. 2004. Quantifying Web-site visits using Web statistics: An extended cybermetrics study. *Online Information Review* 28(3): 191–99.

Kaushik, Avinash. 2007. *Web analytics: An hour a day*. Indianapolis: Wiley Publishing.

Krug, Steve. 2006. *Don't make me think: A common sense approach to Web usability 2nd ed*. Berkeley, CA: New Riders.

Murdock, Theresa. 2002. Revising ready reference sites. *Reference & User Services Quarterly* 42(2): 155–163.

Net Applications, Top browser version share trend for May, 2006 to April, 2007. http://marketshare.hitslink.com/report.aspx?qprid=7 (accessed May 7, 2007).

Nicholas, David, Janet Homewood, and Paul Huntington. 2003. Assessing used content across five digital health information services using transaction log files. *Journal of Information Science* 29(6): 499–515.

Nicholas, David, and Paul Huntington. 2003. Micro-mining and segmented log file analysis: A method for enriching the data yield from Internet log files. *Journal of Information Science* 29(5): 391–404.

Nicholas, David, Paul Huntington, and Peter E. Williams. 2002. Evaluating metrics for comparing the use of Web sites: A case study of two consumer health Web sites. *Journal of Information Science* 28(1): 63–75.

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

Nicholas, David, Paul Huntington and Nat Lievesley. 2000. Evaluating consumer Website logs: A case study of the Times/The Sunday Times Website. *Journal of Information Science* 26(6): 399–411.

Nicholas, David, Paul Huntington, Nat Lievesley, and Richard Withey. 1999. Cracking the code: Web log analysis. *Online & CD-Rom Review* 23(5): 263–69.

Novotny, Eric. 2004. I don't think I click: A protocol analysis study of use of a library online catalog in the Internet age. *College & Research Libraries* 65(6): 525–37.

Pesch, Oliver. 2004. Usage statistics: Taking E-metrics to the next level. *The Serials Librarian* 46(1/2): 143–54.

Peters, T. A. 1993. The history and development of transaction log analysis. *Library Hi Tech* 11(2): 41–66.

Phippen, A. D. 2004. An evaluative methodology for virtual communities using Web analytics. *Campus-Wide Information Systems* 21(5): 179–84.

Sterne, Jim. 2002. *Web metrics: Proven methods for measuring Web site success*. New York: Wiley.

Welch, Jeane M. 2005. Who says we're not busy? Library Web page usage as a measure of public service activity. *Reference Services Review* 33(4): 371–79.