8-20-2002

When Vendor Statistics Are Not Enough: Determining Use of Electronic Databases

Amy S. Van Epps
Purdue University, vanepa@purdue.edu

When Vendor Statistics Are Not Enough: Determining Use of Electronic Databases

Amy S. Van Epps

Abstract – Many libraries have large collections of electronically available databases including journal article and conference paper indexes, full-text vendor catalogs, and standards databases. Which of these resources are being used and to what level becomes a point of interest. A quick re-direct Web-log has been created to track the number of times a particular link is selected, providing a consistent comparison of different resources. The resulting information can be used to determine if what the library provided is being used and if it can be marketed more effectively, which, ultimately, will aid in a cost/benefit analysis for budget decisions.

Keywords – database statistics, scientific and technical libraries statistics, database use, collection development statistics

Amy S. Van Epps holds a Bachelor of Arts in Mechanical Engineering from Lafayette College, an M.S.L.S. from the Catholic University of America, and a Master of Engineering in Industrial and Management Engineering from Rensselaer Polytechnic Institute. Amy is currently Assistant Engineering Librarian and Assistant Professor of Library Science at Purdue University and maintains an active membership in the American Society for Engineering Education, Engineering Libraries Division.

When Vendor Statistics Are Not Enough: Determining Use of Electronic Databases

As library resources for all disciplines become increasingly available in electronic format, libraries are faced with many new issues. Among these are changing demands for library and information instruction to students, marketing of library services and determining which of the many electronic resources our clientele is using.

Being able to determine which resources are being used, and at what level, is the first step in addressing questions such as: Are the electronic materials the library supplies being used? Is the library providing access to what is needed? Do some resources need to be promoted more vigorously? Additionally, librarians need to know who is using the resources, and if users are connecting from an on-campus computer in a given school or department or from a computer in an off-campus location. As Dowling states, librarians need to quantify information since the ability to do so may determine continued funding for the resources (Dowling 2001). In times of tight budgets, libraries are often faced, at best, with flat materials budgets while prices for journals and online resources continue to rise at percentages larger than inflation. Use data should help determine which items are most critical to maintain. At first consideration, the fact that these resources are being supplied and used online would lead one to believe that gathering and analyzing data on which resources are being used most frequently would be relatively simple.

Unfortunately a number of obstacles arise in this seemingly simple task. In many instances, vendors of the electronic products provide statistics on use but only at their own discretion. Multiple vendors lead to multiple formats and different statistical data. If a library receives all of its electronic materials through one vendor, then it is likely to have very few problems, but most science and technology libraries do not have that option. In order to provide all the critical resources, libraries must use a variety of vendors and the

statistics from those various vendors are not consistent or easily reconciled with each other, or in some cases, are non-existent. One vendor may report the number of searches performed, another the number of sets returned, a third the number of records retrieved and/or the number of records viewed, while a fourth may provide the number of logins to the file. Furthermore, when one vendor provides multiple files, e.g. Cambridge Scientific Abstracts (CSA), the number of logins to the vendor service is provided, and each file provided by that vendor shows the number of queries performed. Even if two vendors report the number of searches, the data may not be comparable, as each vendor may have a different definition of what constitutes a search and when a new one begins. Covey identifies incompatibility of data along with multiple formats, delivery methods and schedules for providing data, and the lack of intelligible data as usual complaints with vendor supplied statistics (Covey 2002).

Several organizations are formulating standards for vendor reporting of electronic product statistics. A first stop toward addressing these problems was taken by the International Coalition for Library Consortia (ICOLC), which wrote guidelines for the minimum information that vendors should be supplying to libraries (ICOLC 1998). The guidelines include a list of expected data elements including: number of queries (and a definition of a query), number of session/logins as a measure of simultaneous use, and the number of turnaways, if applicable. More recently, the Association of Research Libraries (ARL) is sponsoring an E-Metrics study on developing statistics and performance measures for electronic materials. Phase One of this project identified current practices for statistics and performance measures in ARL libraries using surveys and site visits, and organized a group to begin talking with vendors about statistics (Shim 2000). Phase Two

of the ARL E-metrics project has gone on to define a number of recommended statistics for library networked resources and assist in defining how these statistics could be used (Shim 2001). The goal of these groups is ultimately to gain agreement and compliance by the vendors on a consistent set of statistics, thus relieving libraries from gathering and analyzing their own data to gain the information needed to make decisions, and to assist libraries in putting their statistical information to use.

The problem of irreconcilable statistics generated the current project at Purdue University. Like several of the libraries in the ARL study, Purdue needs a concrete indication of the most used resources. While the library receives information from most vendors and generates numbers for items loaded locally, the numbers provided are most helpful in tracking use trends for a particular file, not in comparing the resources to each other.

A pilot project of gathering statistics was started to provide numbers that, while not perfect, can be compared to each other in a meaningful fashion and provide data on which resources are being used. The information is created by placing a redirect script call before the URL for each resource. The information gathered is sometimes referred to as a 'click-through'. The redirect CGI script writes a line to a log file that records which resource link was followed before passing the user along to the requested resource. This log files is then analyzed to create the information regarding which of the files are used most often.

Transaction log analysis, which is the technology being used, is by no means new (it has been in existence for about 25 years) (Covey 2002), nor is Purdue the first library to apply it. Transaction log analysis is the process used by most library management systems to determine the types of searches performed most often in the library catalog and other

electronic resources. The ARL E-metrics project visited the University of Pennsylvania libraries and learned they use a similar system to track what they call 'attempted logins' (Shim 2001) and at Texas A&M a click-through page is displayed when a user selects an electronic journal from the library catalog to count the number of times the journal is used (Burford 2001). Using an actual intermediary page may be helpful to display license or copyright agreements for electronic journals, but adds an unnecessary step if your goal is to count click-throughs. Scripts, like the one in Figure 1, provided the ability to track use by writing to the log and automatically redirecting the user to the resource requested.

```
$curlog = "redirect-cgi";
 $delim = "\n";
 $field_sep = "\t";

 ($sec, $min, $hour, $mday, $mon, $year) = localtime( time );
 $mday = '0' . $mday if (length( $mday ) < 2);
 $TimeOnly = sprintf("%02d:%02d:%02d", $hour, $min, $sec);
 $month = (Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov,Dec)[$mon];
 $yr = 1900+$year;
 $DateOnly = $mday."/".$month."/".$yr;
 $tzone = "-0500";
 $logdate ="[".$DateOnly.":".$TimeOnly." ".$tzone."]";
 &re_direct( "$ENV{'QUERY_STRING'}" );
 open( FILE, ">>$curlog" );
 print FILE "$ENV{'REMOTE_ADDR'} - $ENV{'HTTP_REFERER'} - logdate
\"GET".$ENV
 {'QUERY_STRING'} . " HTTP/1.0\" 301
 1\n";
 close( FILE );
 sub re_direct {
 local ($location) = @_;
 print <<"--end--";
 Content-type: text/html
 Location: $location

 <h1>301 Redirect</h1>
 Document is located at <a href="$location">$location</a>
 --end--
 }
```

FIGURE 1
Redirect Logging Script

There are several shortcomings to this type of data collection. First, since it is still in a limited implementation at Purdue, we are not tracking all the use of our files, only those uses that are originating from the Engineering Library's Web page. A comparison of the logged redirect numbers to the total number of institutional logins supplied to us by the vendor indicates a large number of uses are originating from pages other than on the Engineering Library's Website. That is, a large number of users have bookmarked the vendor Web site and are going directly to the site or starting from a page other than the Engineering Library Web page. Second, there is no guarantee that each click-through is an actual use of the database. For example, when a librarian is showing a patron where to click and what to expect, adding a line to the log file, that indicates a different type of use, as well as when people follow a link in error. Finally, researchers who use a particular resource on a regular basis are likely to have that file bookmarked, therefore bypassing our logging process altogether and going directly to the resource. Despite all the places where errors can be introduced into the data, there are many positives to the numbers being gathered. The data is consistent enough for meaningful comparison, and include users information not available from vendors.

Libraries are likely facing an upcoming budget crunch for support of electronic products similar to that experienced for journal subscriptions in the early to mid-1990s. Most librarians, whether or not directly involved with collection development, are aware of libraries experiencing journal reductions in the past. As budgets become tight again, many institutions are facing these issues, only this time electronic resources are also being scrutinized. Before a money problem arises, data is needed to understand what resources are being used, and perhaps learn why some of the important files are underutilized. The

numbers can help libraries initiate conversations with users about their needs and where they are being met, and also help us know what files should be marketed and to which audiences. As with any collection development decision, the numbers will help support a choice, but would never be the sole factor in deciding resource reductions. Research, curricular and institutional needs are among the criteria for evaluation and ultimate decision.

The study began with a request from the Management and Economics Library at Purdue University to gather click-through numbers on their electronic abstract and index resources. Initial analysis was done with a free program called *Analog* that analyzes Web log files on a PC running almost any operating system. The software was chosen for its ability to run on a Windows NT machine and the opportunity for the analyst to specify the format of the log file to be analyzed. The output from Analog is an HTML file that shows the most requested URLs, ranked highest to lowest, and how many times during a given period they were requested. The time period presented is determined by the dates included in the log file. For example, if the log file is divided into one-month segments, then each analysis provides information for use during that month. Other information included the heaviest use times of day and days of the week. This analysis provided the primary information desired.

After working with the information for the Management Library, the Engineering Library staff recognized the applicability and asked to be included in the tracking. A short addition - the CGI script call mentioned earlier - was made to the Engineering Libraries links for those resources to be tracked. Adding a piece to each URL is not as overwhelming as it may sound, provided the library uses a Web-site management program.

In such programs, it is possible to change a link in one location and have all occurrences of that URL change throughout the site. The log file analysis, now being done by the Libraries' Information Technology Department rather than the Engineering Library, uses a package called *WebTrends*®. The software can output the analysis information in an Excel spreadsheet, allowing additional manipulation of the information. Another advantage of *WebTrends*® is the ability to designate the specificity of analysis on the originating IP addresses for each request.

Figure 1 shows the brief Perl script that creates the log file and redirects the user to the requested resource. Analysis of the log file generates the statistics. Since the nature of the logging is minimal, transferring a very small amount of information each time it writes to the log, the user does not see any lag time before connecting to the resource to be used.

Figure 2 includes sample lines from the log file. The sample shows the information tracked, which includes the IP address of the requesting machine, the originating URL, the date and time of the request, the URL that is being requested, and the protocol being used. The last numbers represent the status of the request and the number of bytes sent in response to the request (Dowling 2001). The client information, which includes IP address and referring page, is freely available information gathered from the Web browser as part of the HTTP protocol.

```
168.229.4.1 - http://thorplus.lib.purdue.edu/engr/civil.html -
[01/Oct/2001:10:11:33 -0500] "GET http://hwwilsonweb.com/ HTTP/1.0" 301 1
128.210.124.40 - http://www.lib.purdue.edu/engr/trindexes.html -
[01/Oct/2001:10:37:27 -0500] "GET http://212.49.195.109/webCD/CGI.EXE
HTTP/1.0" 301 1
```

FIGURE 2
Sample Lines from the Log File

The first run of data included April through mid-September 2001 and provided general numbers which presented largely what was expected. Reviewing the report generated by *WebTrends*® showed that our most requested resources were those that provide the full-text of vendor catalogs and standards from IHS. A study done in 1986 showed that practicing engineers used product catalogs as their primary information resource (Jones 1986). At this time the URL for the full-text standards and the product catalogs is the same, so future work will include separating these two resources to determine which is creating the highest use. Looking at the vendor-supplied information on the number of logins indicates that the majority of the use for these resources is being generated by the online standards, nearly three times the product catalog use for this time period. It would be interesting to see the 1986 study redone in light of more materials being available online. Second, by order of most requested items, are the primary engineering databases, Compendex® (Engineering Index online) and INSPEC®. This finding is also in line with the 1986 study by Jones and LeBold. The uses of these files are being logged separately through unique URLs, but *WebTrends*® is truncating the URL requested before the unique part of the string is read. As a result, it is not possible to determine if Compendex® or INSPEC® is used more often. Further refining of the *WebTrends*® profile should address this problem.

Why is this data gathering of interest now? Constant budget questions drive a librarian's desire to quantify where libraries get the "biggest bang for our buck" and which databases could potentially be dropped if the money was not available. The need for comparative statistics, which cannot be determined from vendor supplied data at this point, created the need to develop a process of our own. The use of a redirect log to count click-

throughs fills the data need at a basic level and provides user information unavailable elsewhere. This data gathering provides a good starting place for learning which of the resources the library provides are being used and how often that use occurs.

References

Burford, Nancy and Heather Goetz. 2001. Tracking E-Journal Use from the OPAC. *Poster session presented at the Voyager User Group Meeting*, Des Plaines, IL, April 20-21.

Covey, Denise Troll. 2002. Usage Studies of Electronic Resources. Chap. 3 in *Usage and Usability Assessment: Library Practices and Concerns.* Washington, D.C.: Digital Library Federation, Council on Library and Information Resources.

Dowling, Thomas. 2001. Lies, Damned Lies and Web Logs. *School Library Journal:netConnect* (Spring):34-35.

International Coalition of Library Consortia. 1998. *Guidelines for statistical measures of usage of web-based indexed, abstracted, and full-text resources*. <http://www.library.yale.edu/consortia/webstats.html>.

Jones, R, and WK Lebold. 1986. Keeping Up to Date and Solving Problems in Engineering. In *Proceedings, 1986 World Conference on Continuing Engineering Education*. Lake Buena Vista, FL.

Wonsik "Jeff" Shim, Charles R. McClure, and John Carlo Bertot. 2000. *ARL E-Metrics Project: Developing Statistics and Performance Measures to Describe Electronic Information Services and Resources for ARL Libraries: Phase One Report.* Tallahassee, FL: Information Use Management and Policy Institute, School of Information Studies, Florida State University. <http://www.arl.org/stats/newmeas/emetrics/phaseone.pdf>.

Wonsik "Jeff" Shim et al. 2001. *Measures and Statistics for Research Library Networked Services: Procedures and Issues. ARL E-metrics: Phase II Report.* Tallahassee, FL: Information Use Management and Policy Institute, School of Information Studies, Florida State University. <http://www.arl.org/stats/newmeas/emetrics/phasetwo.pdf>.