Information Use Management and Policy Institute
School of Information Studies, Florida State University

Strategies for Benchmarking Usage of Electronic Resources Across Publishers and Vendors

By

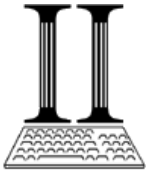
Wonsik "Jeff" Shim, Ph. D. wshim@fsu.edu
Assistant Director, Information Institute, and
Assistant Professor

For

Association of Research Libraries

DRAFT

October 8, 2004



School of Information Studies
Information Use Management and Policy Institute
Florida State University
Tallahassee, Florida 32306
http://www.ii.fsu.edu

Strategies for Benchmarking Usage of Electronic Resources Across Publishers and Vendors

By

Wonsik "Jeff" Shim

Background

The prevalence of electronic contents (e.g., scholarly journals, magazines, books and other reference materials) enabled librarians to gather various indicators of how their collections are being used. However, the data from library's web server logs and usage reports from database vendors proved to be difficult to collect consistently and compare because of the lack of standardization in terms of data elements, definitions and data formats.

The ARL (Association of Research Libraries) E-Metrics Project is an ongoing effort for describing and measuring electronic resources having its roots it the early work on expenditures for electronic resources done by ARL throughout the 1990s.[1] From April 2000 and December 2001 additional research was funded by ARL and conducted through the Information Use Management and Policy Institute[2] to establish additional measures for statistical compilation among ARL member libraries.

The E-Metrics project is one of several major initiatives that specifically deal with measurement of the character and nature as well as use of electronic contents.[3] Other initiatives include works by International Standards Organization (ISO), National Information Standards Organization (NISO), International Coalition of Library Consortia (ICOLC), and COUNTER (Counting Online Usage of NeTworked Electronic Resources) that encompass various kinds of organizations that have stake in the matter. For example, Project COUNTER in particular is working towards standardizing usage reports by establishing several levels of the Code of Practice (COP). ARL was one of the founding members of Project COUNTER and is currently supporting Project COUNTER as part of the ARL E-Metrics set of activities. ARL has also had

[1] Association of Research Libraries, *ARL Supplementary Statistics* (Washington, DC: Association of Research Libraries, annual)

[2] Rush Miller, Sherrie Schmidt, Charles R. McClure, Wonsik "Jeff" Shim, and John Carlo Berot, ~~[multiple authors]~~ *Measures for Electronic Resources (E-Metrics) Measures for Electronic Resources (E-Metrics)* (Washington, DC: Association of Research Libraries, 2002.

[3] Julia Blixrud and Martha Kyrillidou, "E-Metrics: Next Steps for Measuring Electronic Resources" *ARL Bimonthly Report 230/231* (October/December 2003)

discussions with OCLC Office of Research and Marketing staff about the feasibility of further study and standardization in the area of usage statistics by publishers and database vendors.

One of the major findings from the earlier E-Metrics studies was that it takes a significant amount of staff time and effort to gather usage data from database vendors mainly because of its distributed nature. As one librarian put it, "I have too much plumbing to do and have little time for data analysis, let alone interpretation of data." In other words, unless there is a better mechanism to collect usage statistics from more than several different places (or vendors), libraries are not able to make informed decisions in a number of critical areas including resource allocation, system design and user training. The current situation cannot and should not continue. Individual libraries need better, practical tools and approaches not only to be able to collect usage data consistently but also utilize data in a meaningful way to support decision-making.

There have been a number of compilations over the last several years regarding usage statistics of electronic contents. Among them the first survey of the topic was written by Judy Luther in 2000 that was commissioned by the CLIR (Council on Library and Information Resources). The second white paper was produced in 2004 by the Association of American Publishers (AAP) under the editorship of Bernard Rous. It solicited contributions not only from a number of individuals from the vendor side but also from people who represent libraries, users, authors and industry standards initiatives. As a result, it portrays the issue of generating and using usage statistics in a more balanced angle. This paper attempts to not only establish major accomplishments over the years but also suggest future directions especially from the point of view of libraries and library organizations. Therefore, major emphasis on the evaluation of current statistics reporting practices and possible remedies to improving the collection and use of vendor usage statistics reports.

What is usage statistics?

In the context of electronic subscription-based databases, usage statistics usually refer to the indicators of the volume of user access to electronic resources and services available from content providers or vendors (Shim & McClure, 2002). Examples of those indicators are a count of sessions is a specific database, the time per session in a specific database, the count of searches performed, and the number of times full-text documents are downloaded. At the most basic level, they are "nothing more than the artifacts of online transactions (Rous, 2004)." They are akin to a transaction record for a gas station that will show how many gallons of gas a customer purchased, the time of transaction, and the amount of money paid and so on. However, unlike the transaction at the gas station, the online transaction leaves several important questions still answered: Was

the item (an electronic document or its surrogate) really what the person was looking for? What is the purpose of the query? What is the value of the item to the user? Because of these caveats, caution would be necessary to interpret usage data.

Source of data

Currently most content providers supply some sort of usage reports to their customers. We summarize some of the trends in vendor usage reports in the section entitled Trends in Usage Statistics Reporting. But libraries themselves also cull usage statistics, some using web traffic analysis software such as *WebTrends* and *Analog* and others using in-house custom scripts and relational database. Nowadays, many libraries are employing a click-through method to collect usage statistics internally. Using a click-through mechanism, libraries can direct user access to licensed databases to an intermediate web page that serves as an electronic counter. The kinds of information that are usually captured include: number of times the page is accessed, the destination (database), Internet address of the requesting machine, time/date and so on. The mechanism works only for database access through library web pages. If the user bypasses the library web site (e.g., typing the database vendor's website directly or through stored bookmark), that access cannot be captured. The big advantage of the click-through mechanism is that the uniform usage data can be collected by the library as opposed to obtaining inconsistent usage data from different vendors.

Table 1 summarizes pros and cons of vendor reports and library internal reports.

Table 1: Comparison of vendor and library generated usage reports

	Advantages	Disadvantages
Vendor Reports	▪ Comprehensive ▪ Takes less effort to collect	▪ Difficult to compare ▪ Takes time and effort to collect statistics from multiple vendors
Library Reports	▪ Data are more consistent ▪ Libraries have control over data	▪ Not comprehensive ▪ Takes time to build the technical infrastructure

It is reasonable to believe that there is a relatively high correlation between usage reported externally by content providers and internally by libraries themselves, unless there is a marked difference in the way that libraries provide access to electronic materials.[4]

[4] A preliminary statistical analysis done at the University of Pennsylvania Libraries shows evidence supporting the claim.

Factors affect the usage numbers

We need to be aware of a host of factors that affect and/or skew usage numbers. Some of the well-known factors that are believed to influence usage level are:

- How the numbers are captured and counted: Vendors have different technical infrastructure and different philosophies of counting. An obvious example is how vendors apply time-out parameters to compute session counts. In the case of aggregate databases, is a search to multiple database packages counted as a single search or as a separate search for each database chosen? Is browsing a secondary database such as author, subject, or journal list counted as a search or a menu selection? Does the vendor take into consideration multiple clicks for the same item in a short period of time (for example, less than 10 seconds) and treat them as one request or multiple requests? Is clicking the next button to retrieve the next set of results counted as separate search? What kind of filtering is in place to weed out non-user access? For these reasons, many librarians feared that collating usage statistics from multiple vendors is like comparing apples and oranges. Fortunately, initiatives such as the Project COUNTER and the ICOLC guidelines are having tremendous positive impact on standardizing the counting method.

- Usability of library web page: It is a fact that web page interface, that is how the information is structured and presented on the digital space, affects use. Providing multiple access points (A-Z list and by subject), labeling resources with clear and helpful descriptions, helping users with remote access all contribute increased use. Electronic resource management application package such as *Serials Solutions* and *TDnet* are becoming important tools to effectively present electronic materials to end users. In addition, linking services such as *SFX* (from Ex Libris), *Article Linker* (from Serials Solutions) and *icate* (from Openly Informatics) significantly enhance user navigation among various electronic information sources regardless of different vendor platforms and interfaces.

- User population: Other things being equal the size of users or the size of the research community for a given field is an important factor that affects the level of use of information contents. Therefore, in a way, high usage of certain research titles does not necessarily indicate their importance or value relative to other titles (Rous, 2004).

- Instruction and marketing: Library instruction that focuses on specific electronic database(s) will yield increased use, for at least a short period of time after the instruction. More importantly, effective marketing of electronic resources to targeted user groups will positively affect use of those resources.

Uses of usage statistics

One tends to think that libraries are the main consumer of usage statistics. But libraries are not the only group that has vested interest in usage statistics. Content providers probably are keener on usage statistics for a number of reasons. Consortia administrators would like to examine usage statistics across member institutions to make sure that the funding is adequately distributed based on usage. In this section, we describe various parties that use usage statistics and describe in general terms how they use the statistics.

Content Providers

Granted different content providers have different views and motivations for usage statistics (e.g., aggregator vs. publishers), first and foremost, they collect usage statistics to support and to validate their business model. For example, content aggregators such as Ebsco, Gale and ProQuest has been collecting usage statistics so that they could calculate royalties to publishers.[5] Link out counts—user access directed to the publisher website or gateway from the aggregator website for full-text— are increasingly sought after by publishers. Usage statistics are also sought to analyze content, package products, improve delivery systems and fine tune system capacity (Pesch, 2004). When there is a seat limit (known as simultaneous access quota), vendors use the turnaway statistic, that is the number of times user access was denied because of insufficient number of seats), to encourage the library to increase the quota. Vendors would look at both high usage and low usage by the libraries and develop strategies for contract negotiations that take into consideration of usage level. More and more, the provision usage statistics is becoming a standard practice as part of official contract requirement or of customer service.

Libraries

[5] Normally the royalties are based on the amount of content rather than actual use. However, publishers would like to find out how the level of demand on their contents. Likewise, aggregators use usage data to assess the effectiveness of contents from multiple providers.

As expenditures increasingly shift from print collection to electronic materials, managers of libraries at all levels are seeking ways to assess the use of electronic resources. Usage statistics have become important because of their role in support of analysis, reporting and decision-making (Shim, Murphy and Brunning, 2004).

The most highlighted use of usage statistics is when librarians need to make subscription decisions, i.e., cancellation and renewal of journals and databases, although usage is not the only factor that affects the decision. Librarians need usage statistics to assess user behaviors in and demand for electronic contents, help identify strengths and weaknesses of different electronic resources, and design access mechanism to better support user demands. Libraries also need statistics to fulfill various reporting requirements that support accountability, record keeping, justification for funding, and strategic planning. Underlying the analysis and reporting functions aforementioned is often the need for information in support of specific decisions: resource allocation, system design and configuration, funding and support, and resource sharing (e.g., consortium relationships). The extent to which libraries are using usage statistics is not publicly known. Allegedly, somewhere between 20-50% of libraries are actually using vendor reports based on access to vendor usage report system.[6]

Consortia

People who manage various library consortia or resource sharing arrangements such as state-wide licensing agreements will want to obtain usage reports from their vendors to justify the expenditures and assess products they offer to member libraries. If the funding comes from the state or local government, it is customary that usage statistics are required as part of the justification process. In addition, usage reports inform these managers of the effectiveness of their portfolios for collection management and of any areas of improvement for training and system design. In many cases, usage reports for individual member libraries of a consortium are available for all the other members for comparison and benchmarking. In those circumstances, the consortium will also take the responsibility of making usage reports of all participating members available through its website.

[6] These were the figures mentioned by vendor representatives at the Network Performance Measures Task Force Meeting and the NFAIS forum on online usage statistics. We believe that the figures are somewhat inflated due to the fact that many libraries opt to have scheduled reports sent automatically via email when they become available.

Users as Readers and Authors

One would not imagine users of electronic materials as consumers of usage of those materials. However, authors whose writings included in the electronic subscription collection would show a lot of interest in getting usage statistics of their writings. At least one publisher[7] allows its authors to access usage statistics that include monthly online usage (abstract, full-text HTML and PDF requests), citation information, and comparisons to similar articles (Anderson, 2004). Tenopir and Baker (2004) advocate that access to statistics about certain characteristics of the journal or article, and the local subscription may help users to make judgments. In the current climate, end users who are accustomed to rankings and popularity indexes provided by Internet websites such as Amazon.com may well embrace these statistics. On the other hand, the negative impact of including those indicators needs to be weighed against the potential benefits.

Evolution of usage statistics

Over the last several years, we have seen a surge of various initiatives from library and standards organizations to standardize data elements and reporting requirements. These initiatives are different in a number of different dimensions.[8] Each initiative is unique in terms of its organizational nature (e.g., formal standards organization vs. informal body of libraries and library consortia), its primary interest (e.g., statistical reporting, reporting practices for negotiating contracts, and standardizing report format and delivery), and by its scope of interest (e.g., entire library services vs. electronic resources). Major initiatives among them are:

- NISO (National Information Standards Organization) Standard: In 2004, NISO updated its Z39.7—*Library Statistics Standard*—to include network services and resources statistics and performance measures. (see http://www.niso.org/emetrics/current/complete.html) and is out for comment at the present time. This standard extends beyond the December 2001 ICOLC, ARL E-Metrics Project, ISO, and IMLS definitions. Members from ARL, ICOLC, NCLIS, public libraries, special libraries, and vendors participated in the development process.

[7] The New England Journal of Medicine.

[8] Based on David Goodman's presentation during the NFAIS forum on online usage statistics on October 1, 2004 in New York. (Presentation available at http://www.nfais.org/events/event_details.cfm?id=26)

- ICOLC (International Coalition of Library Consortia) Guidelines for Statistical Measures of Usage of Web-based Information Resources: ICOLC published the first guidelines in November 1998 and subsequently revised it in December 2001 (see http://www.library.yale.edu/consortia/2001webstats.htm). Regarded by many as the de facto standard in usage reporting before Project COUNTER came along.

- ARL (Association of Research Libraries) E-Metrics Project: The project was a concerted effort by twenty-four member libraries to develop and test statistics to describe electronic resources and services. Project reports including data collection manual are available at http://www.arl.org/stats/newmeas/emetrics/contract00-01.html. Based on the project recommendations, ARL will be collecting E-Metrics measures during the 2003-2004 period (July 2004) through Supplemental Statistics Program. In addition, measures related to expenditures for electronic resources are now moved to the main ARL statistics.

- ISO (International Organization for Standardization): Two working groups under the auspices of Technical Committee (TC) 46 are currently reviewing library statistics (ISO 2789) and performance indicators (ISO 11620), both of which related to network statistics.

- IMLS (Institute of Museum and Library Services): As a U.S. federal agency, IMLS sponsored three projects to develop national network statistics and performance measures for public libraries and state library agencies. The Information Institute conducted these studies to develop public library network statistics, build a national data collection system, and conduct librarian, state library agency staff, and consortia staff training. Reports and various documents from these projects are available at http://www.ii.fsu.edu.

- NCLIS (National Commission on Libraries and Information Science) Task Force on Network Performance Measures: NCLIS has continued its work in standardizing online database usage statistics and reporting mechanisms. The taskforce, which has completed its fifth meeting and is now run by the Information Use Management and Policy Institute, is quickly becoming a major venue for people involved in various initiatives to exchange information on each other's work.

- EQUINOX: Funded by the European Commission, EQUINOX is a project to address challenges in performance evaluation in increasingly networked library environment. Specifically it attempts to develop performance measures for electronic library and to design and test an integrated information system to store and analyze the measures. However, there has not been direct influence from EQUINOX on many of the initiatives listed above. Project reports are available at http://equinox.dcu.ie.

- DLF (Digital Library Federation) ERM (Electronic Resource Management) Initiative: Released in August 2004, the report from the ERM initiative provides a roadmap for systems that are designed to manage a variety types of electronic resources by providing functional requirements, specifying data elements and establishing relationships among those elements. While not directly related to usage statistics per se, people can clearly see its potential impact on usage statistics reporting. People who were involved in the ERM design recognized the work of the Project COUNTER and intentionally left the usage statistics part of the design up for COUNTER standards. So once the COUNTER standards are firmly in place, the usage statistics portion of the ERM will be plugged by the COUNTER standards.

Project COUNTER

Most recently, a group of publishers, intermediaries, and library organizations have formed a consortium called COUNTER (Counting Online Usage of NeTworked Electronic Resources). COUNTER's aim is to develop and maintain a Code of Practice (COP) that governs the measurement and reporting of the use of online information products and services. COUNTER released its first draft of the COP in December 2002 with the draft of Release 2 pending publishing in January 2005. Figure 1 summarizes the required reports to be compliant with COUNTER through Release 1 and Figure 2 shows realigned COUNTER reports in draft release 2.

Figure 1: Mapping of required data elements in COUNTER COP (Release 1)

	Journal Report	Database Report
Sessions		Level 1
Searches	Level 2	Level 1
Full-text Articles Requests	Level 1	
Item Requests (Page Types)	Level 2	
Turn-aways	Level 1 (Full-text requests) / Level 2 (Item requests)	Level 1

*Level 1 report is mandatory to become COUNTER compliant. Level 2 is recommended.

Figure 2: COUNTER COP Reports (Draft Release 2)

Journal Report 1:	Number of Successful Full-Text Article Requests by Month and Journal
Journal Report 1a:	Number of Successful Full-text Requests in html and PDF Formats
Journal Report 2:	Turnaways by Month and Journal
Database Report 1:	Total Searches and Sessions by Month and Database
Database Report 2:	Turnaways by Month and Database
Database Report 3:	Total Searches and Sessions by Month and Service

The COUNTER COP also specifies the definitions of reported data elements, report formats, data availability, report access, as well as guidelines as to how vendors process data to ensure consistency in the way statistics are generated.

COUNTER is significant because 1) it is a collaboration of vendor and library communities, 2) it is international in scope, and 3) it applies to all types of libraries. While we don't expect Project COUNTER to solve all of the problems associated with collecting usage statistics once and for all, it may well prove to be a cornerstone towards an environment where libraries are able to systemically collect useful and reliable usage statistics. As of September 2004, COUNTER has more than 150 members and 32 compliant vendors.

Two major developments at Project COUNTER are worth further comments. One has to do with auditing process. Currently vendors who wish to declare compliance can do so by providing a signed declaration. Starting in 2005, however, annual external compliance review becomes mandatory for vendors to earn the COUNTER compliant designation. COUNTER has completed the development of the auditing procedure with the requirement expected to be implemented in January. The auditing will be conducted by a professional financial auditor or by a qualified person approved by COUNTER. The specific auditing rules are not yet known. But they will likely to involve a script based validation where a separate account is set up for each vendor for testing purposes and matching is conducted between the statistics generated by following the script by the auditor and the statistics reported by the vendor in addition to checking various report format and delivery requirements. Allegedly, a variance of –8% and 2% between the script generated statistics and the vendor reported statistics is being discussed.[9]

Another development is the draft release of an XML DTD (eXtensible Markup Language Document Type Definition)[10] based on the COUNTER Release 1 reports. The specification, developed by Timo Hannay of the Nature Publishing, is a first step towards an environment

[9] According to the presentation by Oliver Pesch at Ebsco during the NFAIS forum on online usage statistics held on October 1, 2004 in New York. Mr. Pesch is a member of the COUNTER Executive Committee.

[10] Unlike HTML (HyperText Markup Language) which defines formatting of web pages, XML is used to define and organize contents.

where vendors allow external software agents to selectively gather usage data elements from their data servers or surrogates.[11] While the publication of the XML DTD opens up the possibility of exploring the feasibility of establishing data harvesting mechanism, it will take some time for the vendors to organize their data in XML format and for the libraries to possess the capability to retrieve XML based data made available by vendors.

In summary, the standardization of usage of subscription-based electronic contents has seen substantial improvement over the last several years. Critical to this success was the fact that individual initiative has been well informed of other related initiatives and the fact that libraries and content providers had realistic expectations for each other.

Trends in usage statistics reporting

As part of the E-Metrics project, in the summer of 2000, we conducted analysis of usage reports from 10 vendors in terms of data elements included, their breakdown by journal or database title or by IP address, and report format and delivery. What we found was that despite all 10 vendors provided some data, they are not detailed enough (that means provide only aggregate data), limited option for report format (in many cases only HTML tables which are difficult to process), and non-existent documentation. All in all, vendors were just beginning to understand the needs of libraries and to build primitive system for providing usage reports. However, at the same time vendors and library representatives continued conversation using regular venues such as NCLIS network performance measure task force meetings in conjunction with ALA mid winter meetings.

Four years after we conducted our initial investigation, we went back and looked at usage reports to see if there has been any improvement. In the meantime, two very important guidelines related to the standardization of usage reports were produced. The ICOLC guidelines for usage reports came out of library consortia group. The COUNTER Code of practice is collaboration between vendor and library communities. In terms of the usage statistics enumerated in the two standards, there is a significant overlap. One major difference between them is the fact that the ICOLC guidelines include the number of menu selections (e.g., use of alphabetic and subject menus) to capture different modes of user behaviors in the system. Overall, the COUNTER COP puts more emphasis on the formatting of usage reports and makes distinctions between reports from database vendors and full-text journal providers.

We compared the usage reports from 18 vendors—8 content aggregators and 10 full-text journal providers—in terms of data elements provided and report format and delivery. One thing we would

[11] The technique is commonly referred to as data harvesting.

like to mention is that this is not a scorecard of usage statistics. The fact a vendor received more yes's does not necessarily mean that the vendor provides qualitatively better report. The comparison shows only a snapshot of essential elements in usage reports for overall comparison. Also the selection of vendors is fairly arbitrary: the fact that a vendor is included in the comparison does not indicate the quality of its report.

Table 2 summarizes key data elements provided by the select vendors. Note that there is a varying degree of data available. The table combines the data elements from the ICOLC guidelines and the COUNT COP. In terms of granularity, aggregators report by database packages or individual files whereas full-text journal publishers report at the journal title level with the only exception of IEEE Explorer in which only summary statistics are reported.

Table 2: Comparison of key data elements from select publishers and vendors (as of Sept. 2004)

	Sessions (Logins)	Queries (Searches)	Menu Selections	Full Content Units Accessed	Turn-Aways
★ **Content Aggregators**					
Gale InfoTrac	Yes	Yes		Yes	Yes
EBSCOhost	Yes	Yes		Yes	n/a
Lexis-Nexis		Yes		Yes	n/a
OCLC(FirstSearch)	Yes	Yes		Yes	Yes
Ovid	Yes	Yes		Yes	n/a
ProQuest		Yes	Yes	Yes	n/a
SilverPlatter	Yes	Yes		Yes	Yes
Web of Science	Yes	Yes		n/a	n/a
★ **Full-text Journal Providers**					
Blackwell Synergy	Yes	Yes	Yes	Yes	n/a
HighWire		Yes	Yes	Yes	n/a
IEEE Explorer				Yes	Yes
Ingenta Select	Yes	Yes		Yes	n/a
JSTOR		Yes	Yes	Yes	n/a
Kluwer		Yes	Yes	Yes	Yes
Oxford Univ. Press		Yes	Yes	Yes	n/a
Project MUSE	Yes	Yes		Yes	n/a
ScienceDirect	Yes	Yes	Yes	Yes	n/a
Wiley InterScience	Yes	Yes	Yes	Yes	Yes

* n/a: not applicable or not available.

Menu selections, according to the ICOLC guidelines, refer to the use of various menu systems (e.g., alphabetic and subject menus) that are provided by the vendors as alternatives to direct searching. The table shows that full-text journal providers are more likely to report menu selections probably because their contents are more amenable to browsing than the contents provided by the aggregators.

Table 3 summarizes various aspects of usage statistics report delivery as well as the COUNTER compliance status of the reports. The ICOLC guidelines specify that usage reports be accessible through an interactive webpage and separate from the administrative functions that

deal with subscription management. With the only exception of JSTOR report[12], all of the vendors in the comparison provide access to usage reports through a password controlled website. The table also shows that the same user id and password are being used to access not only usage reports but also licensing management areas for many vendors. This can potentially be an issue when several people share access to usage reports page.

Table 3: Comparison of vendor reports in terms of report delivery (as of Sept. 2004)

	Password controlled website	Separate login for usage reports	Report format(s)	Delivery Options	COUNTER Compliance
★ **Content Aggregators**					
Gale InfoTrac	Yes	No	HTML, Text	Email	Yes
EBSCOhost	Yes	No	HTML, text	Email	Yes
Lexis-Nexis	Yes	Yes	HTML, Text	Download	No
OCLC(FirstSearch)	Yes	Yes	HTML	Display only	No
Ovid	Yes	Yes	Text	Display only	No
ProQuest	Yes	No	HTML, Text	Email	Yes
SilverPlatter	Yes	Yes	HTML, Text	Email	No
Web of Science	Yes	Yes	HTML, Text	Download	Yes
★ **Full-text Journal Providers**					
Blackwell Synergy	Yes	No	HTML, Text	Email	Yes
HighWire	Yes	No	HTML, Text, Excel	Download	Yes
IEEE Explorer	Yes	Yes	HTML, Text	Download	No
Ingenta Select	Yes	No	Flash, Text	Email only	Yes
JSTOR	No	n/a	HTML, Text	Display only	No
Kluwer	Yes	Yes	HTML, Text	Download	Yes
Oxford Univ. Press	Yes	Yes	HTML, Text	Download	Yes
Project MUSE	Yes	Yes	HTML, EXCEL	Download	Yes
ScienceDirect	Yes	Yes	HTML, Text, Excel	Download	Yes
Wiley InterScience	Yes	No	HTML, .Text	Download	Yes

* Delivery option includes screen display unless otherwise noted.

Overall, we have the following observations from the comparison.

- First, in terms of the amount of data and detail, there has been a significant improvement. All of the reports analyzed have proper breakdown of data by journal or database title, by month or by week, and so on. All 18 reports were being accessed through the web. We had several reports in our previous analysis available through print report or by email request only. Most reports were available in multiple formats, usually HTML for screen display and some form of text file (most likely CSV, comma separated format) for downloading. Several vendors offer 6-7 different formats with the option for file compression.

[12] The JSTOR usage report is available from any legitimate IP addresses from which access to the JSTOR content can be made. So, the web access to the JSTOR usage report page from the Institution A will display usage report for Institution A. Theoretically anybody who uses the legitimate IP address can access his or her institution's usage report for JSTOR.

- We were very impressed with the level of data and report delivery from individual vendors. There have been significant improvements on the individual vendor level. One of the things that librarians realized through the discussion with vendor representatives was that it takes a significant commitment of resources to develop usage reporting modules. Apparently many vendors made the investment and began providing some useful data.

- The fact that 12 out of 18 major vendors in the comparison are COUNTER compliant with respect to their reports is a testimony to the COUNTER accomplishment over a relatively short period of time. In particular, when we conducted a similar comparison in August 2003, only three vendors were COUNTER compliant showing an accelerated buy-in from the vendors. Since the COUNTER audit is not yet implemented, we don't know what will happen to the COUNTER compliance status for these vendors. However, we expect that the majority of the vendors will attain COUNTER compliance this time in 2005.

  There is a couple of remaining issues that need further comments.

- While terminologies and report formats are being converged thanks to initiatives such as ICOLC guidelines and COUNTER code of practice, reliability and validity of data still remain an issue. By reliability, libraries would like to see consistent counting from the same vendor over time and across vendors. There have been several anecdotal evidences that suggest what libraries are collecting internally and through usage reports were counting are not the same as what usage reports say. In this area, we do hope that the work of COUNTER narrows the gap through its planned auditing process.

- Another important development was that as the amount of data increases, so does the complexity of system interface. The amount of data you can retrieve from some vendors such as Elsevier's ScienceDirect and Kluwer can easily drown many uninitiated people.

- While several vendors now provide some form of documentation, we really want to encourage all vendors to do more in this area. Perhaps they need to include a step-by-step instruction on how to download and process data files. Also more information is needed for librarians to better understand how the numbers are generated and processed.

- We emphasized that there has been significant improvements from individual vendors. But libraries do not deal with one or two vendors. Academic libraries and many public libraries deal with several dozen vendors and accordingly several dozen different reports and usage reporting systems. It takes an enormous amount of time and effort to collect these reports and put them in a usable, standardized format for analysis and decision making. Improving access through a tightly standardized report format is one way to improve the situation. A central place or repository for usage report information can also be extremely beneficial and needs serious consideration.

There are also several interesting developments that might interest some. One is the concept of fee for service. Ingenta Select charges $250 annually for advanced statistics package. Basic statistics are provided free of charge. OCLC has similar set up for interlibrary loan statistics. Right now, Ingenta Select appears to be the only vendor who charges for statistics but we may see more of this.[13] Second, two vendors outsourced usage reporting to data warehouse vendors. Kluwer's reporting system is hosted by IBM and ScienceDirect was using Ordina (MicroStrategy is the application name used by the vendor) for supplying usage data. Not coincidentally both vendors were the ones with the largest amount data. While professional service may improve the quality of data and access, too much complication can also work against what most users, librarians, ask for.

Issues affecting effective collection and use of usage statistics

There are a number of technological, environmental, logistical, and interpretational issues and challenges that affect the collection and use of vendor usage statistics. We will describe these factors in this section.

Logistical Factor

In late 1990's and early 2000's when vendors had just begun offering rudimentary usage statistics and libraries had begun asking for usage reports, the kinds of concerns we heard coming from librarians were: We are getting the data, The data are neither consistent nor comparable. And the data are not detailed enough to be useful. Now that most vendors provide usage statistics, some too much data, the concerns and complaints are now shifted to the sheer logistics of

[13] We asked the participants of the Network Performance Measures Task Force Meeting during the 2004 ALA Mid-Winter in San Diego about this possibility. The general consensus was that usage statistics would remain a basic service and free of charge in the foreseeable future.

collecting usage reports from multiple vendors. Even with the advancing standards such as the Project COUNTER and the ICOLC guidelines, the distributed nature of vendor usage reports and inconsistencies of the way the reports are organized and delivered, collecting usage reports has become a very daunting information management task.

While most academic and public libraries have at least one full-time staff member that deals exclusively with electronic resources, the licensing part of electronic materials is what gets the bulk of their attention. While collecting usage statistics seems to be an integral part of their job, they are often left with insufficient time, tools and training that would enable them to stay on top of tracking usage of electronic materials. Let us give a reality check here. The electronic resources librarian at a medium size state university keeps a MS Word file that lists the webpages where vendor reports are available and user id and passwords to those webpages. The document is 9 pages long and maintaining it is in and of itself a lot of work.[14] It is easy to imagine how much work that needs to go into obtaining the reports, putting them into a local server and massaging the data to a level that other staff can use. Judged individually, vendor usage reports seem pretty palatable. But when combined, as is the case for the majority of libraries, they become beyond manageable. The situation inevitably leads to very limited and ad-hoc use, if not non-use, of vendor usage reports.

Technological Factors

No doubt technologies alter the conditions in which various statistics are captured. Three specific examples of new and emerging technologies are worth mentioning: federated search, linking and Shibboleth.

Federated searching[15] is a relatively new but rapidly growing technology. It allows users of electronic contents to use a common interface to search and retrieve from a myriad of electronic contents provided by the library. In other words, it is a meta search engine applied to library databases. Federated search, or search portal, raises questions regarding how things like searches and sessions are to be counted for reporting purposes. Unless meta searches are captured separately, for example by redirecting them to a separate IP address or by relegating them to a separate id, both session and search counts will be artificially bloated. Currently NISO has established the Metasearch Initiative that "will identify, develop, and frame the standards and

[14] Some vendors require changing passwords every month to make the situation even worse.

[15] Also known as meta searching, parallel searching, broadcast searching or cross-database search. See http://www.loc.gov/catdir/lcpaig/portalproducts.html for a listing of existing federated searching products on the market.

other common understandings that are needed to enable an efficient and robust information environment"[16]. We hope that the initiative provides guidance as to how to handle meta search activities in usage reports.

Linking through the standard called openURL allows a user who found a citation for an article or a full-text item to access the appropriate copy of the object, be it at the publisher site or at a gateway service. The process involves a link resolver service that picks up the preference arrangement for the institutional user with the content provider without the express acknowledgement by the user. The linking is not necessarily a roadblock to measurement. However, it engenders significant changes in the way users go about navigating and retrieving electronic materials. Also, currently linking activity—liking out from abstracting and indexing services to publishers—is not specified in either COUNTER COP or the ICOLC guidelines, although several aggregators have begun reporting link-out activities.

Shibboleth is an Internet2 standard that allows transparent authentication for external resources. Applied to subscription database access, it has the potential for a user to use single institutional id (e.g., campus user id or an employee id) to access distributed electronic materials based on the pre-established access criteria for the user between the institution and the content providers. At the moment, the technology is not being implemented by any of the major content providers as most of them grant user access by matching the requester's IP address against the legitimate IP address pool defined for the user's institution. However, as more and more educational contents cross campus boundaries, it is not unreasonable to expect that libraries, education institutions, and content providers will adopt the technology in the future. If implemented, libraries will be able to gather a cleaner data at the individual user level that has long been sought but that has not been recorded. The flip side of this is the resulting liability for privacy and confidentiality infringements.

Environmental Factors

The most important transformation that is taking place in the realm of scholarly publishing should be the one related to open access (OA). [17] Every day, there is a discussion, a proposal, a forum that discuss the shape and characteristics of open access models. Whether the open access model will overcome the "serial crisis" or whether it will become the dominant method of

[16] http://www.niso.org/committees/MS_initiative.html

[17] To keep up with current debates on open access, refer to the Nature magazine's discussion focus at http://www.nature.com/nature/focus/accessdebate/.

disseminating scholarly works is yet to be played out. The fact of the matter is open access is already occurring. The challenge from open access movement for usage statistics is not so much about whether they will be methods by which usage statistics can be collected from distributed access points for OA articles, but as the need for and the use of usage statistics of OA. In other words, in the current subscription based model, libraries need usage statistics primarily to make subscription decisions. In the OA model, since libraries are responsible for neither the publishing nor access costs, they will not be the primary consumer of usage statistics or usage estimates (Harnad, 2004). Therefore, we need to rethink the nature of usage statistics from a very different perspective in the pre-dominant OA environment.

Interpretational Factors

Usage statistics of electronic materials, whether there are generated internally or obtained from content providers, pose a number of interpretational challenges. First of all, due to insufficient experience in the electronic environment and due to the lack of comparable statistics from the print environment, it is very difficult to establish context for understanding the data currently available (Luther, 2001). The same sentiment is echoed by Rous (2004), "The first problem in understanding any usage statistics for online literature is that there is no baseline (p.7)." Clearly a non-use means something. But it is often difficult to determine how much is adequate.

Librarians would like to compare their usage of electronic contents with peer institutions that have similar collections. Certainly the statistical data compiled by library organizations such as ALA and ARL will be of use for comparison purposes. However, because of the level of granularity of data reported to these organizations, the aggregate statistics will not permit the kinds of comparison that individual libraries would like to conduct.

Cost per journal article accessed is often used to determine cost effectiveness of a journal or the whole deal. But it is an incomplete measure in the sense that the number of articles included in individual journals varies journal by journal. Many librarians are eager to receive training and guidance in the form of best practices workshop, online tutorials, and so on.[18]

Due to the fact that usage statistics by their very nature are quantitative, interpretation of the figures and the justification made from the interpretation assume that high or increasing usage is always good. But sooner or later, some of these statistics will hit a plateau and the numbers will

[18] The Information Use Management and Policy Institute at Florida State University has developed a web-based instructional system related to e-metrics. Available at http://www.ii.fsu.edu/emis/.

stagnate. In light of this, librarians will need to look for alternative or qualitative measures that can speak for the value of electronic materials and services to the user community.

Recommendations

First of all, for the libraries, we recommend that they need to determine how much data they want to collect and how they will be using the collected statistics. The amount of effort it takes to collect necessary information can quickly escalate and libraries will need to get organized and have realistic expectations about how much data they can process and use. The standardization of usage reports and delivery, most notably by the Project COUNTER, will help immensely in terms of cutting down the amount of time and effort to retrieve statistics. However, ultimately usage statistics need to be processed, used, interpreted and presented and that is entirely the library's responsibility.[19] Libraries should not wait until the standards organizations take care of data definitions, report format and the level of data.

In terms of specific recommendation that can allow us to move forward in collecting and using vendor usage statistics, we strongly believe that the time has come to seriously think about the need for a central depository for usage statistics. In the remaining section, we provide general descriptions of the idea of central depository to spur specific discussions and proposals.

Central Depository for Vendor Usage Reports

Need

Despite the good work that these projects were able to accomplish and will continue to do so, there are serious roadblocks to making use of usage statistics from an individual library's point of view, such as the amount of work needed to collect and process these statistics and the lack of resources at the individual library level. These roadblocks are key to understanding why many libraries are not using usage statistics. Certainly, vendors need to provide standardized data elements according to standard definitions in an agreed upon format. However, that appears to be a long-term goal that can only be accomplished by libraries, vendors and government agencies working together in a concerted effort. Libraries are looking for ways in which to retrieve and analyze the data with minimum effort NOW.

[19] Refer to Shim, W. (forthcoming). Getting the Most out of Vendor Supplied Usage Data. In J.C. Bertot, & D.M. Davis (Eds.), Planning and Evaluating Library Networked Services and Resources. Westport, CT: Libraries Unlimited.

The results of the ARL E-Metrics clearly indicate that for libraries to be able to collect, use and report data, they need to invest heavily in infrastructure and personnel that can support data acquisition and processing activities. Currently, many libraries do these manually, especially with usage statistics from database vendors. Although it appears that the usage statistics webpages at vendor websites are being widely used to disseminate usage statistics to libraries, librarians still need to visit all of these websites—a number usually running up to several dozens—and download reports, typically on a monthly basis. Another problem is that the reports come in different formats (e.g., spreadsheet, text, or HTML format) to make it difficult or impossible to combine these reports into a coherent dataset suitable for institutional analysis. Due to these and other problems, a significant number of librarians do not attempt to process the reports and store them for further analysis.

The current practice is that libraries update their summary of usage statistics of vendor databases—mostly likely in a spreadsheet format—and the raw data are thrown away. Some libraries do keep vendor reports without converting them in a standardized format because it is a very labor-intensive process. In this environment, librarians rely on quick spot checks to find out any oddities in the data. This situation does not permit libraries to conduct systematic analysis of data to find out underlying usage patterns and other valuable information. Overall, there is a huge price to pay, both up front and on a continuing basis, if a library is serious about employing electronic materials usage data.

Some libraries have begun sharing information and specific ways to deal with obtaining and massaging vendor usage statistics through mailing lists such as liblicense (library license hosted by Yale Library) and ERIL (Electronic Resources in Libraries). However, a single library is not likely to commit significant resources to embark on an initiative to remedy the situation. What we need is an independent organization—neither a library nor a vendor—to serve as a portal for data retrieval and other crucial services to foster the use and analysis of usage statistics.

Key Features of the Proposed Depository

The depository will consist of a web interface and accompanying software solutions that can expedite the retrieval of usage statistics from database vendors. The proposed solution will create an environment in which libraries collect vendor usage statistics with minimum effort, save them into a standardized data format, and conduct various data analysis—all within familiar, standard web browsers. The solution will allow libraries to perform report generation on multiple vendor reports and journal-title level data extraction and usage pattern analysis.

The web-based interface should be flexible enough to allow libraries to retrieve customized reports of core usage statistics based on the COUNTER standards. Users of the interface will be able to specify a time period, database vendor, and journal (or database) title from which to retrieve data. Instead of

maintaining several dozen user names and passwords for different databases to access and retrieve vendor specific usage data, librarians will use one profile to conveniently access and manipulate data in one place. The web-based portal eliminates the most labor-intensive process—the actual acquisition and processing usage statistics—allowing librarians and decision makers to focus on data analysis and interpretation.

The key functions of the proposed system will include (See Appendix A for a graphical view of key processes):

- Acquiring usage statistics (including past statistics) directly from vendors;
- Organizing imported statistics from multiple vendors into a standardized format on the trusted server hosted by an independent body to be identified (most vendor specific data elements will be kept);
- Displaying both summary (e.g., by product) and detailed (e.g., by journal or database title) statistics over the Web using a single, integrated interface;
- Exporting data in standardized formats (e.g., database, spreadsheet, text files) for storage and archival purpose;
- Performing comparative analysis across institutions and by vendors (individual libraries cannot look up usage statistics of other libraries). The institutional comparison will be based on library size, institution type (e.g., public vs. private), and other factors. Currently, JSTOR is the only vendor who provides comparative usage statistics;
- Collecting and distributing key documentations, examples, and software tools related to usage statistics reporting and analysis;
- Maintaining up-to-date information of main contacts for usage statistics at database vendors; and
- Providing a forum (discussion board or mailing list) that can help establish a community around the issues of usage statistics reporting.

Most individual libraries do not have the necessary resources or expertise to develop such a system. More importantly, vendors are most likely unwilling to respond to conflicting demands from a host of libraries. If libraries can agree on a single system of sign-on and access, it will make the job much easier for and attractive to vendors.

Libraries can continue to access usage statistics from individual vendor websites. However, the web portal provides a necessary separation between the functions of administering institutional profiles, which should be done at the vendor websites, and accessing usage statistics through the web portal, which can be shared among several librarians responsible for collection development and data

analysis. There are several initiatives that attempt to establish technical standards for facilitating the management of electronic resources using open architecture such as XML (extensible Markup Language). DLF (Digital Library Federation) ERM (Electronic Resource Management) Initiative is one such effort. Dealing specifically with usage statistics, the Project COUNTER released the draft XML DTD for transmitting vendor usage statistics in an automated fashion. The depository effort will need to work closely with these initiatives in the discussion and establishment of such a standard and align system development and maintenance efforts along the emerging standard.

Benefits

The long-term solution to the effective usage statistics reporting and utilization by libraries hinges on a number of factors including standardization of data elements, definitions, and reporting format. However, the proposed solution will eliminate the main obstacle for acquiring usage statistics and greatly enhance the retrieval of usage statistics. We believe that the proposed solution will facilitate actions and research toward the long-term solution. In addition to the main benefit, we expect the following outcomes from the establishment of a central depository of vendor usage statistics:

- Libraries will be able to benchmark their usage levels with those of peer institutions at a more granular level that cannot be achieved by the aggregate statistics published by library associations. By centrally hosting the datasets, this would become a reality.
- Vendor documentations regarding definitions of reported statistics and how they are calculated will become available centrally, allowing for cross comparison.
- Librarians will be able to exchange information on usage statistics more effectively.
- It will foster the exchange of examples of analyses and toolkits used at individual libraries.
- It will eliminate the need to hunt down vendor contacts since the depository will maintain and update the information.
- The depository will serve as a testbed for continued research on the standardization of usage reporting. It will also contribute to the coordination of various initiatives in this area such as the ICOLC guidelines, COUNTER Project and others.
- By contributing data to the depository, database vendors can improve their service to subscribing libraries. Also, vendors will have a central channel to disseminate information regarding their usage statistics reporting.

Issues and Challenges

There are a host of issues that need to be resolved for the central depository to be established and provide the kinds of service that it promises. One important issue has to do with who should be the independent third part that will host the service. Clearly we will need to address this question. Also, we need a sound business model for the depository. The hosting and application development will require both monetary and personnel resources. Another important issue has to do with the privacy and confidentiality of personal information that arises because of the transmission of usage statistics from vendors to the depository. One of the privacy principles that has evolved over the years deals with onward transfer in which "a data collector who transfers data to a third party should ensure that the third party provides at least the same level of privacy protection as does the data collector itself (p.138)." (Marks & Lange, 2004). These are not trivial issues and they all need serious considerations.

Conclusion

As libraries expand their offering of external resources and services for their users, the importance of usage statistics from outside content providers will grow. Therefore, usage statistics are becoming a crucial component for determining the effectiveness of library resources and for assessing the outcome of library services.

The proposed depository is a specific and practical solution that fulfills these needs. The complexity and the magnitude of the problem confronted by many libraries would require more systematic and coordinated approach to achieving more efficient and efficient assessment of library services and their impact.

References

Anderson, K.R. (2004) Usage data for authors. In B. Rous (Ed.). Online usage statistics: A publisher's guide. New York: Association of American Publishers.

Association of Research Libraries. ARL Supplementary Statistics. Washington, DC: Association of Research Libraries. Annual.

Blixrud, J., & Kyrillidou, M. (December 2003). E-Metrics: Next steps for measuring electronic resources. ARL Bimonthly Report, (230/231).

Franklin, B., & Plum, T. (2004). Library usage patterns in the electronic information environment. Information Research, 9(4). Available at http://informationr.net/ir/9-4/paper187.html.

Harnad, S. (2004, Apr. 12). Re: Rethinking "collections" and selection in the post Gutenberg age. Message posted to the American Scientist Open Access Forum, archived at http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/3702.html.

ICOLC (International Coalition of Library Consortia) guidelines for statistical measures of usage of web-based information resources. Available at: http://www.library.yale.edu/consortia/2001webstats.htm

Institute for the Future (2002). Final synthesis report for the e-journal user study. Stanford University Libraries. Available at http://ejust.stanford.edu/research_findings.html.

Ke, H., Kwakkelaar, R., Tai, Y., & Chen, L. (2002). Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's SienceDirect OnSite in Taiwan. Library & Information Science Research, 24, 265-291.

Luther, J. (2001). White paper on electronic journal usage statistics (2nd ed.). Washington, D.C.: Council on Library and Information Resources.

Marks, B.E., & Lange, D.M. (2004) Usage statistics, Internet privacy and the right to read anonymously. In B. Rous (Ed.). Online usage statistics: A publisher's guide. New York: Association of American Publishers.

Miller, R., Schmidt, S., McClure, C.R., Shim, W., & Berot, J.C. (2002) Measures for Electronic Resources (E-Metrics) Measures for Electronic Resources (E-Metrics). Washington, DC: Association of Research Libraries.

Pesch, O. (2004) Usage statistics: The full-text aggregator. In B. Rous (Ed.). Online usage statistics: A publisher's guide. New York: Association of American Publishers.

Rous, B. (Ed.). (2004). Online usage statistics: A publisher's guide. New York: Association of American Publishers.

Rous, B. (2004) Introduction: Use and abuse of online statistics: Overview. In B. Rous (Ed.). Online Usage Statistics: A Publisher's Guide. New York: Association of American Publishers.

Shim, W. (forthcoming). Getting the most out of vendor supplied usage data. In J.C. Bertot, & D.M. Davis (Eds.), Planning and evaluating library networked services and resources. Westport, CT: Libraries Unlimited.

Shim, W., Murphy, K., & Brunning, D. (2004) Usage statistics for electronic services and resources: A library perspective. In B. Rous (Ed.). Online usage statistics: A publisher's guide. New York: Association of American Publishers.

Shim, W., & McClure, C.R. (2002) Improving database vendors' usage statistics reporting through collaboration between libraries and vendors. College & Research Libraries, 63(6), 499-514.

Tenopir, C., & Baker, G. (2004) Usage statistics for users as readers, authors, and collection development stakeholders. In B. Rous (Ed.). Online usage statistics: A publisher's guide. New York: Association of American Publishers.

Appendix A:
Simplified Diagram of Key Processes
for the Central Depository for Vendor Usage Statistics