Where does it come from? Generating and collecting online usage data: an update

The proliferation of standards and methods for collecting and reporting usage statistics for online resources has begun to turn any discussion of such into a type of alphabet soup. Additionally, the ever developing relationships between libraries, vendors, and now to a much greater extent, developers and sundry other 'IT people' has further confused the conversation surrounding the effective use and reporting of these statistics. This article will attempt to provide a 'big picture' view of the generation and collection of usage data from both technical and practical perspectives.

TANSY MATTHEWS
Associate Director, Virtual Library of Virginia
George Mason University

A bit of background

Usage statistics are not a product of the electronic age. Now, as always, usage data tracks patron activities that libraries have long recorded – information seeking and resource use. In the physical library, these are interpreted as gate counts, reference questions and re-shelving activity. Online, the same behaviors may be translated as sessions, queries and full-text downloads. In both environments, it is important to note that the data captures what may be called 'symptoms' of intellectual activity. A patron may enter the library and do no research, just as an online user may initiate a session and walk away. A user may take a bound journal from the shelf and never open it, just as a user may download articles and not read them. As librarians, however, we accept the symptoms of use as indicative of the ways in which the library collections are being used.

The ongoing increase in the amount of research being conducted online and the proliferation of electronic journals and e-books provides new opportunities for understanding the ways in which patrons are utilizing our collections. Rather than assuming that a journal taken from the shelf has been used, we can now know how many individual articles are being viewed. The basic units of print publication (journal, book) are breaking down into articles and chapters, allowing us an increasingly granular understanding of what is being used and how. At the same time, however, libraries are increasingly dependent on vendors to report usage data, and vendors are under increasing pressure from libraries to report standardized data in a timely manner. Where libraries have been recording intellectual activity in the forms of gate counts and re-shelving for decades, the identification and quantification of these activities is still a fairly new challenge for publishers and other information vendors.

How is usage data generated?

A vendor's challenge, then, is to identify 'symptoms' of these intellectual activities in an online environment, recognize them, and count them. Given that the 'symptoms' in an online environment are similar to those in the print realm (i.e. a request for information, seen through an article download or an unshelved journal, indicates that information-seeking activity is taking place) there are two means by which this can be accomplished. The first is by using a web server's inherent technology – the web server log. A web server log maintains a

history of the requests put to the server. However, in a web server log, an entry is created every time a link is clicked or an image is downloaded; this would capture, for example, not only full-text downloads, but a line might be created every time the vendor's logo was accessed to appear in a web page. Much of the data contained in a web server log is thus of little use, and extracting data from web server logs is impractical for most vendors.

As an alternative, most vendors' systems include a locally created internal logging infrastructure that does not rely on server logs. These infrastructures can be constructed to count only the transactions the vendor wishes to record; for example, database searches and deliveries of full-text data. Initially, this led to a problem for librarians in using the statistics that were provided – as each vendor created their own logging system, different things were being counted, and different definitions were being applied to the same terms. Librarians needed to know that they were receiving similar data across platforms, otherwise the numbers for each vendor existed in a vacuum and could not be considered in relation to each other. The solution to this problem emerged in Project COUNTER, a project initiated in 2003 as a co-operative effort between libraries and publishers to standardize the reporting of use data (http://www.projectcounter.org).

What role does Project COUNTER play in how vendors generate statistics?

The Project COUNTER Codes of Practice (COP) provide 'an agreed international set of standards and protocols governing the recording and exchange of online usage data.'[1] The COP specifies: the data elements to be measured; definitions of these data elements; usage report content, format, frequency and methods of delivery; protocols for combining usage reports from direct use and from use via intermediaries.[2] Though it does provide guidelines for data processing by vendors, the COP does not specify *how* the data should be collected; the decision on how to log usage is left up to the individual vendor. Therefore, vendors that wish to become COUNTER compliant are free to work within their own systems to extract the specified data elements and format them appropriately for delivery. This allows for maximum flexibility on the part of COUNTER participants – as long as the data that is being reported complies with Project COUNTER's guidelines, the information provider's underlying system may need to be altered very little.

What challenges do vendors face in becoming COUNTER compliant?

As explained by Oliver Pesch, EBSCO Chief Strategist of E-Resources and member of the COUNTER Executive Board, the challenge vendors face in implementing COUNTER varies greatly depending on the collection system currently in place. For example, Release 3 of the COUNTER COP (published August 2008) includes a new protocol that requires 'federated searches and automated searches to be isolated from *bona fide* searches by genuine users.' Translated, this means that robot (i.e. spidering) activities will need to be excluded as well as activity generated by federated search engines. The changes represent distinctly different levels of challenge to vendors. The former, excluding robot activity, may not prove to be difficult for most vendors. If the vendor is processing usage statistics from a web server log, excluding this data is simply a matter of matching the recorded browser ID (a standard entry on a web server log) to a crawler and deleting the activity. If the vendor is using an internal framework that does not currently record the browser ID, the logging format will need to be changed to include the extra element. While both of these changes will have to wait for an opening in the development queue, the new protocol is unlikely to require major changes to a vendor's system.

The accompanying requirement, isolating federated search activity, represents a potentially more difficult challenge and is known to be impossible in some vendors' current systems. For example, vendors that capture search data without capturing session information will not be able to break out information on federated searches – they simply do not have the data to do so. These vendors may be unable to comply with the updated COP without undertaking a fairly major system redevelopment.

In the case of larger vendors, making these changes may be a matter of waiting for their next system release or development cycle. For smaller vendors or publishers, the problems of collecting and reporting statistics in general, and compliance

in particular, may be increased by a lack of technical or development staff. As a result, many smaller vendors and publishers are moving to third-party vendors for statistics provision and site maintenance. Given the expectations on vendors for meeting standards such as COUNTER and archiving systems such as Portico[3] and LOCKSS[4], it has become far more feasible for smaller vendors to contract out to a third party for these services.

What software is involved in collection and reporting of usage data?

As discussed above, on the vendor side the collection and reporting of usage data is entirely up to the particular vendor. Third parties such as HighWire Press, Scholarly IQ, and SurfAid Analytics provide these services on behalf of many smaller vendors. On the library side, however, there are products commercially available to assist libraries in collecting usage data from a variety of vendors, either as stand-alone applications or as part of a larger umbrella product. Scholarly Stats (Swets), 360 COUNTER (Serials Solutions), and Journal Use Reports (Thompson Reuters) are all stand-alone products. Ex Libris and Innovative Interfaces, Inc. both include usage statistics collection features in their electronic resource management systems (ERMs).

In all cases, the products rely on COUNTER-compliant reports provided by vendors as a basis for a variety of cost-per-use and usage level/ranking reports. Intended to facilitate usage collection for libraries, these services do not require highly technical implementations on the library side – those that are not contained within an ERM are accessed through web interfaces. An important common denominator in these products is that they currently rely on, or plan to implement, the SUSHI protocol to automate collection of usage data, making the consolidation of data viable across a large number of both libraries and vendors.

The Standardized Usage Harvesting Initiative (SUSHI) Protocol is a NISO standard that 'defines an automated request and response model for the harvesting of electronic resource usage data utilizing a Web services framework.'[5] In plain terms, SUSHI makes it possible to automate the retrieval of COUNTER-compliant usage reports by providing a standard 'wrapper' around COUNTER XML files. This means that, to a computer system, the data will always look the same, regardless of the vendor it comes from. A programmer can then develop a service that uses a single 'call' to get the data – if the data was not standard, a different call would have to be used for each vendor, an unfeasible prospect.

For institutions that do not subscribe to a product, like an ERM, that can 'grab' their statistical data, SUSHI offers the possibility of an open-source solution. Working with Adam Chandler and others, Tommy Barker at the University of Pennsylvania has developed a web service that can place the 'call' and retrieve the XML data from the vendor's server. He is currently working on a client toolkit that a programmer will be able to use to read the data that the vendor's server returns. Barker warns that the implementation of the service he created is work for a developer with the appropriate knowledge – yet another area where librarians will need to work closely with IT staff in search of a solution.[6]

Hopefully, at some point an open-source 'plug and play' solution will be available. Meanwhile, libraries that do not have the luxury of such technical expertise (my consortium, the Virtual Library of Virginia, falls into this category) will have to decide whether to continue processing statistics manually (as we do) or to invest in a service like those noted above. In our case, we have developed a local system, based on Microsoft Excel and a few 'pirated' macros, for automating the processing of COUNTER-compliant statistics to a point where we are well able to manage the processing in-house and will continue to do so for the foreseeable future. It must be noted, however, that as a consortium that does not share a common integrated library system (ILS), we faced specific challenges that merited the 'experimenting' time that was invested in creating our system, and we still look forward to being able to implement a SUSHI-based solution in the future.

Although the intellectual processes involved in information seeking (i.e. search and browse) still closely resemble those employed in a physical library, the revolution that has occurred in information access over the past two decades has been so profound as to require no elaboration here. At this point, it is up to librarians and vendors to work together to identify and assess the ways in which we are serving our patrons in an ever-expanding online environment. Consistent, reliable

usage statistics can help us in this effort, provided all parties understand the limitations of technology and the demands of staff time on both sides.

References

1. Project COUNTER website:
   http://www.projectcounter.org (Accessed 19 January 2009)
2. COUNTER Codes of Practice. See ref.1
3. Portico website:
   http://www.portico.org (Accessed 19 January 2009)
4. LOCKSS website:
   www.lockss.org (Accessed 19 January 2009)
5. http://www.niso.org/workrooms/sushi (Accessed 19 January 2009)
6. Tommy Barker, personal communication, 10 November 2008.

■ **Tansy Matthews, MLIS**
Associate Director, Virtual Library of Virginia
Fenwick Library B222, MS2FL
4400 University Drive
George Mason University
Fairfax, VA 22030-4444, USA
Tel: 703-993-2694
E-mail: tmatthe6@gmu.edu