

Back Talk — Use Statistics: Are they worth it?

Column Editor: **Anthony (Tony) W. Ferguson** (Library Director, University of Hong Kong; Phone: 852 2859 2200; Fax: 852 2858 9420) <ferguson@hkucc.hku.hk>

Introduction

Recently at the one and only **Charleston Conference** I gave a talk about use statistics. **Katina** suggested that I rewrite it for **Back Talk** so here it is:

I recently composed the following letter to **MacMillan Publishing**.

Dear Sir or Madam:

We greatly appreciate the quality of your reference books, their physical appearance, feel, smell, the quality of the work done by their authors and editors, and the prices, while high, are not totally objectionable.

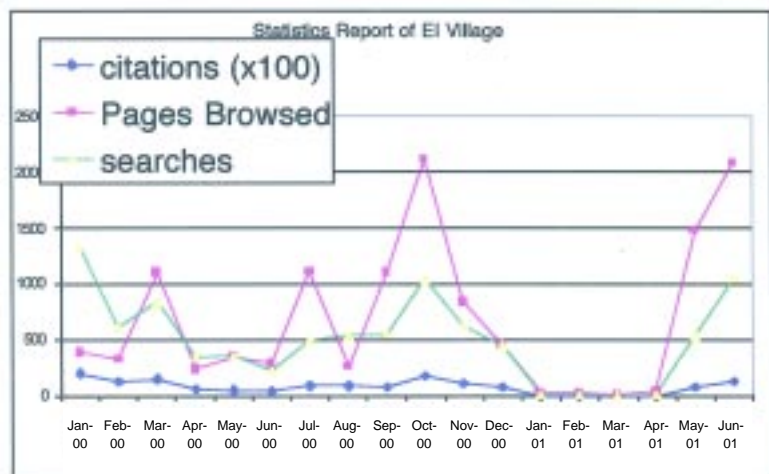
However, we feel that before we can agree to continue purchasing these printed materials, you need to provide them with increased functionality. Specifically, we want you to provide, in addition to the reference books themselves, an automated functionality employing wireless or BlueTooth technology that will record and summarize the way these books are used. Moreover, in line with the ICOLC Guidelines for Statistical Measures of Usage, we want data on how many times your books are opened, the number of times the reader subsequently flips from the table of contents or the index to the main body of the book looking for specific bits of information, and the amount of time spent reading chapters within the book. These bits of information should be analyzed in terms of the reader's demographic characteristics, e.g., major, year in school, sex, etc. Of course, we do want to know the date and time these materials are used with at least monthly summary statistics.

We must alert you to the fact that those publishers who are able to use satellite technology so that this data can be collected wherever the reader might take the volume, e.g., to the beach, bed, bathtub, etc., will also be favored over others. Indeed, if an alarm could be sounded when our book is submerged in water or in the mouth of a dog, such innovative uses of technology will be greatly appreciated and rewarded. Using location honing devices we could then dispatch a rescue team. Please help us preserve our intellectual heritage.

Sincerely Yours,

Now I haven't sent the letter, and, of course, to ask this of print publishers is a bit far fetched — even though with the use of radio frequency information technology at least some of this might be possible. My point is, we are asking a lot of electronic publishers and I believe we need to make sure we want and can use what we have/can get before continuing our quest for digital library use statistics.

At the **University of Hong Kong** we have been keeping track of a variety of statistics for a small subset of the nearly 400 databases currently being offered our patrons. These figures are used to generate **Excel** tables like this one:



It's not pretty data. It's incomplete. But it visually demonstrates why **ICOLC** has provided vendors with a list of data that is wanted. Without uniform data, attempts to provide comparative or even cumulative data are fruitless.

In general, **ICOLC** asks for six bits of information. They want data about the number of sessions, the number of queries, the number of menu selections, the number of articles downloaded, information about who used it (IP addresses or account codes, etc.), breakdowns of the data for the various components of each database, e.g., by title, and they want this information on a regular basis, e.g., monthly. At the **University of Hong Kong** have been able to gather monthly statistics on the number of sessions, queries, and number of full text downloads but not the rest for most of our databases. I believe most universities are able to do similar sorts of things. We are only able to develop an incomplete picture of what is happening.

Why Should We Continue To Gather and Analyze Use Statistics?

Thomas A. Peters in a useful 2002 *New Library World* article summarizes the major reasons why we need to analyze use statistics for electronic forms of information:

- Unlike for print resources, their use is not visible to the eye.
- They are needed to demonstrate to funding agencies what they are getting for their investments.
- Collection developers need the information to decide what to cancel, what to buy less of, and what kinds of products are needed in greater abundance.
- Public service librarians need the data to know what to promote, for which databases might patrons need more help in using, etc.
- License negotiators need data to give them informed leverage for subsequent negotiations.
- Better understand what some old rules of thumb like the 80/20 rule mean in the new digital environment.

Rush Miller, in a recent presentation at the Centennial celebration of the **Peking University Library**, echoed the importance of this data to collection developers. **Linda Mercer** in a Winter 2000 article in *Issues in Science and Technology Librarianship* suggested an additional reason: by analyzing these statistics we can gain an understanding of user preferences for things like PDF vs. HTML.

Peters also summarized the reasons why non-librarians need this data:

- Authors and tenure boards want to know the degree to which article X was read.
- Editors want to know what sorts of articles are read so they can publish more of them.
- Publishers need information to establish how much their product is worth.
- Vendors want to know the best pricing mechanism to use, based upon how their product is already being used.

So Data Is Wonderful, but What Are the Problems Associated with Gathering and Analyzing It?

The literature focusing on the difficulties of gathering and analyzing this data is growing as librarians move from

continued on page 94

the theory to the practice of doing it. Based upon my reading of just a few articles, here are the major issues as I reinterpret them in my own words:

- For librarians, they are very time consuming, expensive to gather (**Peters**)
- For publishers and vendors, they are very time consuming and expensive to produce (**Peters**)
- Analysis designed to decide if your expenses are worth it happen after the fact (**Peters**)
- Many of the ejournal titles you identify to cancel are protected because they are part of packages (**Mercer**)
- Potential for privacy issues to be brought up muddle an already difficult enterprise (**Peters**)
- We are gathering and analyzing reflections of value (use), not information about how valuable these materials are to our readers (**Blecic, et al., Peters**)
- Different publishers don't agree what they mean by sessions, queries, etc., making comparisons impossible and understanding illusionary (**Peters, Miller**)
- Some vendors segment their databases so statistics are doubled when both segments are accessed (**Blecic, et al.**)
- Multiple accesses of the same article in the same session are often counted the same as if multiple articles were downloaded (**Blecic, et al.**)
- Vendors change use-monitoring software/techniques without informing library customers (**Blecic, et al.**)
- Resource time-out functions increase the number of times users have to duplicate their searches to get what they want (**Blecic, et al.**)
- Even when differences between institutions are "normalized for enrollment, major, or credit hour," it turns out that use patterns differ greatly between institutions (**Townly and Murry**)
- Use is determined as much by the prominence a database is given, and the amount of promotion it receives, as its relevance to the needs of a particular user community (**Townly and Murry**)
- Use is a reflection of the length of time a resource has been available, e.g., it takes time for a resource to catch on and be used (**Townly and Murry**)
- Possibly because of these last reasons, publishers and vendors are reluctant to develop and share these statistics (**Ferguson ala Peters**).

The definitions problem is aptly illustrated by a few quotes from **Miller's** Beijing talk which is also published in the May issue of the *Journal of Academic Librarianship* and is based upon meetings with publishers and vendors:

- "One vendor assumes that if the 'search' key is entered more than once within a certain number of seconds, it is an error in that undoubtedly the student or faculty member is double clicking where they should not be and the system automatically adjusts the numbers accordingly."
- "Today the loading of one Web page could well generate 50 hits in the server log because of all the complex elements that are loaded separately to comprise that page."
- "Well then, what is a full text? Is it a single photograph, a paragraph, a caption, an abstract, an entire article, one page of an article or some other element? You might be surprised to learn that (a) almost no two vendors define these things the same way, and (b) few of them provide customers with any definitions at all."

I would add an additional difficulty related to the gathering and analyzing of e-journal use statistics:

- Looking at how much each e-journal is used, just like in the old print journal world, belies the dynamics of what is happening in the new digital world. While an e-journal is still a journal and an electronic article is still an article, these journals and articles are also links in the new network of information made possible by the interlinking of journals. This phenomenon is one of the few really revolutionary changes that have been made possible by the use of computers by libraries (the other major changes are full text searching and the Web itself). When we cut an e-journal we are cutting a link, not just an unwanted debit to our serials or book budget.

So, what should we do, throw up our hands and forget the whole thing?

I believe the literature and common sense indicates there are a number of positive things we can do to move the work along:

- Remember that basically all we are trying to do is to figure out if we are getting sufficient value for our investment. To figure that out we just need to establish value and a use statistic is one such measure, but not the only one. **Rush Miller** put his finger on it when he said: "The real questions we need to answer are not quantitative . . . but are outcomes based. What difference does the use make to our users? Are they learning more in their courses as a result of the availability of digital libraries, are they working smarter, and more efficiently? Are faculty members more productive in scholarship and teaching?"
- As noted by **Peters**, we need to "mainstream" the collection and analysis of use statistics. The simple truth here is that unless someone is assigned to spend time with these statistics for each title, they are worthless.

- Again **Peters** suggests that we need to apply the knowledge we gain through our analysis. It is simply too easy to tell our systems department to develop tables for analysis, or to get mountains of tables from consortia or vendors, and then leave the paper piles on or in filling cabinets.
- If, in our analysis of the imperfect statistics that are available to us, we find that a particular resource seems little used, before cutting it we need to ask ourselves the following questions and take the appropriate actions:
 - Does the target audience easily find the resource? If not, change things.
 - Has the resource been in place long enough to gather an audience? If not, delay action for another six months or a year.
 - Has the target audience been taught to use the resource? If not, figure out the most appropriate user-training program.
 - Does the resource perform a uniquely important linkage role for the target audience? If yes, think twice about cutting it.

None of the above is meant to criticize the ongoing efforts involving librarians, publishers, and vendors to create better statistics, but only to place emphasis on the immediate tasks before all of us as we try to bring our users and the information they need together as quickly and efficiently as possible. At **HKU** we have decided to pull back on our efforts to create homogenous, comparable statistics and, with the exception of internally generated session statistics (which will also be understated), instead to link to publisher statistics for the time being. Once uniform statistics are available, we will go back to generating tables. 🌳

References

Charles, T. Townley and Leigh Murry, "Use-based Criteria for Selecting and Retaining Electronic Information: A Case Study," *Information Technology and Libraries*, v. 18 (March 1999): 32-39.

Deborah D. Blecic, Joan B. Fiscella, and Stephen E. Wiberly, Jr., "The Measurement of Use of Web-based Information Resources: An Early Look at Vendor-supplied Data," *College and Research Libraries*, v. 62 (2001): 434-453.

Linda S. Mercer, "Measuring the Use and Value of Electronic Journals and Books," *Issues in Science and Technology Librarianship*, (Winter 2002) <http://www.library.ucsb.edu/istl/00-winter/article1.html>.

Rush Miller, "Shaping Digital Content," *Journal of Academic Libraries [Beijing]*, (October, 2002), pp. 85-92. Also published in *Journal of Academic Librarianship*, v. 28 (May 2002).

Thomas A. Peters, "What's the Use? The Value of E-Resource Usage Statistics," *New Library World*, v. 103 (2002): 39-47.