December 2003

Usage Statistics at Yale University Library

Jennifer Weintraub

Yale University Library, jennifer.weintraub@yale.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Weintraub, Jennifer (2003) "Usage Statistics at Yale University Library," *Against the Grain*: Vol. 15: Iss. 6, Article 10.

DOI: <https://doi.org/10.7771/2380-176X.4201>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Usage Statistics at Yale University Library

by **Jennifer Weintraub** (Digital Collections Specialist, Yale University Library)
<jennifer.weintraub@yale.edu>

Librarians in the all-print library were happy with their relative understanding of how much their collection was used. They could get information by keeping records of circulation, journal browsing, and even a cursory look at how well worn their books were. In the electronic world, we do not have those cues. At most, without vendor statistics, I can tell how many times someone connected to a vendor database from one page on the library's Website. Usage statistics are essential in helping us become a truly integrated print and electronic library. **Yale University Library** has been collecting usage statistics for electronic resources regularly since 1998. The process has moved from paper faxes to direct access to the statistics, and our interest in the statistics has changed as well. This descriptive article illustrates the kinds of statistics we maintain from vendors, the issues we have in processing and comparing this information, and what we ultimately hope to gain from this work.

We currently collect usage statistics from approximately 50 vendors or publishers, covering over 130 electronic databases, three collections of electronic books, and twelve journal packages that include thousands of journals. This represents about a quarter of our total electronic resources. I am responsible for maintaining these statistics and do so with input from selectors and the help of our Electronic Collections Assistant. At **Yale**, selectors negotiate with vendors with the help of the Head of Electronic Collections and asking for usage statistics for an electronic resource is a regular part of the negotiation process. The selectors then forward the information to me for our domain-restricted Website with our usage statistics or information about how to get access to the material.

We collect these statistics for several reasons. Like many libraries, we monitor our databases for turnaways when we have simultaneous usage, and for usage changes that may signal problems with accessing the database on the vendor

side. Since electronic resource pricing and features fluctuate from vendor to vendor, it is helpful to learn what value we may be receiving from any one specific vendor. **Yale** does not, for the most part, make selection decisions solely based on usage statistics. Finally, we hope to learn more about our users habits in the future through maintaining usage statistics without invading our users' privacy.

We collect statistics on a wide variety of material. There are the easy to define resources, such as electronic journals, electronic reference works (such as the **Oxford English Dictionary**), and indexes to periodical literature. There are also the now familiar but slightly more complicated aggregations, which are frequently periodical databases that are linked to vast quantities of full text (these include the **WilsonWeb** databases, through which we subscribe to resources such as **Art Abstracts** and **Readers Guide Abstracts/Full Text**). In the last two years we have also begun to collect statistics on electronic books. Like most university libraries, our subject and language areas are wide and varied. More importantly for the area of usage statistics, the publishers themselves are at varied levels of technological aptitude, sophistication, and willingness in the area of getting us information about how our students, faculty, and staff are using the electronic resources.

As electronic resources have changed and grown in complexity, our statistics have changed as well, particularly in the amount of information available. Rather than "number of hits" on a Web page I can receive a much greater degree of granularity. I now receive search statistics, browse statistics, most used articles, most used electronic books, what journal publishers were linked to from citations, and number of PDFs viewed, among other interesting facts. However, new problems abound with new kinds of electronic resources and tools. We recently had a question about an aggregator counting links from our OpenURL resolver as searches on particular databases. This led to a high number

of searches on a database rather than views or downloads for a particular journal. The help documentation did not clarify how the statistics were collected (which is more common than you might think). We still don't receive many statistics that show the frequency of cross-database searching (something interesting to us, as we'd like to know what combinations are most used, for example). Some vendors still do not send us anything but "hits" on the database, which is virtually useless, or "views" for an electronic journal without any explanation of what that means. In order for us to truly understand what was going on with that electronic resource, time must be spent to ferret out the answers.

Another area that has changed is presentation or dissemination of the statistics. There are fortunately very few vendors who fax or mail us paper statistics. Increasingly I can now access our statistics via IP authentication or a password that is separate from our administrative password for the vendor. This is our preferred method as I can disseminate this password to our reference librarians and selectors, who can then look for statistics themselves without the possibility that they will change the settings of that database. I also receive information via email in a comma-delimited format and as html files. Despite these technological developments, **Yale** must maintain dozens of passwords to various databases in order to get usage statistics. Collecting usage statistics has become a more complicated facet of electronic resource management than it was when we had 30 databases and four electronic journal packages.

Usage statistics can appear as html, text, or comma delimited files. If the material arrives in html format, I put that Web page up on our server and input some important data (for example, searches per resource) into an Excel file and make a graph as a visual aid for the selectors. When statistics are sent as text, PDF, or CSV files, I make a Web page from the Excel file as well. Thus, most statistics that are sent directly to me eventually become an Excel file. I also make graphs for certain expensive packages if the material, though available on the publisher's own server, is not easy to obtain or read.

Some usage statistics do not even fall under the rubric of "useful" in the most basic sense. Of our available statistics, 60% conform to the ICOLC guidelines. Of the statistics that I can access through our preferred methods, those with IP authentication or non-administrative passwords, however, nearly 70% are satisfactory. And I can get satisfactory statistics for 70% of our most popular databases (the ones described as such on a special Web page). There is a correlation between statistics being easy to get and being high quality. The publisher with the worst statistics is one that claims they will

Rumors from page 30

and aggregators. Topics covered will include publisher policies on article withdrawal, notification of customers, compensation for withdrawn material, and legal aspects of the debate. At least 30 minutes will be reserved for audience comment and discussion, in which digital archiving, treatment of defamatory or inaccurate content, plagiarism, and other topics will be addressed. Presenters include: **Rick Lugg**, R2 Consulting (Moderator), **Davies Menefee**, Director, Library Relations, the Americas, Elsevier, **Sam Brooks**, Senior Vice President, EBSCO, **Dawn Conway**, Vice President, Content & Business Development, LexisNexis.

Stanford University's HighWire Press has launched a new feature for institutions, "**Shop for Journals**." 28 society publishers, with content hosted on **HighWire**, have banded together to create an easy way to select from an initial list of 60 journal titles and create custom packages, with more titles expected to join in the New Year. To support this program, these publishers have developed a standard set of **Guidelines for Institutional Access** (defining authorized use and users), and have agreed to use a common tiered pricing model, based on type of institution. <http://highwire.stanford.edu/shopforjournals> "Shop for Journals" is the gateway to something we are calling an "Open Package," says **Michael Clarke**, Senior Managing

continued on page 40

continued on page 34

send them every quarter and do not. When I do receive them, the electronic journal statistics contain one number: downloads per journal per month. While this is the essential number I want to know, I get no sense how many tables of contents were viewed, whether the package as a whole was searched, much less any linking information. But judging from my cursory survey above, the situation is improving, though still time consuming. I also find most vendors to be receptive to complaints or questions about their statistics and have answered at least five surveys in the last calendar year about electronic resources that had some component devoted to usage statistics.

The inability of even some expensive vendors to provide us with the minimum level of acceptable statistics leads to holes in our analysis of how people are using electronic resources. Yale has been a member of the **ARL E-Metrics** project for the last few years, and as part of this we are asked to test various statistical measures of electronic usage and collections in libraries. The most troublesome question is "Total Number of Searches on Electronic Databases." While it is possible (though incredibly time consuming) for me to give a number of "total searches we are told about," it is unclear what this can really mean to us. It is difficult to extrapolate from our available statistics to what our actual number of database searches were. We also can't tell how different groups are using our electronic resources, or how people are moving from one resource to another, or whether they use our linking tools or are unaware of them. This is important as we assist users in person and implement new services to exploit our full electronic potential.

My future goals for our usage statistics are to increase their relevance to librarians at Yale. First, I continue to strive to get information that can really help them make decisions about what kinds of resources are the best for our users. I also want to be able to demonstrate trends across subject areas, not just vendors. Setting up a "data farm" or database of the statistics is one way we can accomplish this. Finally, we support efforts to get better statistics from vendors and contribute to new research using this data, at Yale, within our consortia, and at the national level.

Our collection of usage statistics, while perhaps not as comprehensive as we'd ideally like it to be, is a step in the direction of being able to tell what our users are doing and what they want and need, even if they can't verbalize it to us. Sitting in our library, it can be hard to tell if the investment Yale made in electronic books was worth the money. Once we created an eBooks database and put it prominently on our Web page, usage of electronic books increased immediately. Certainly it is hard to tell how satisfied people are with those books aside from the fact that usage keeps growing every month. But we do know that people are using them, and perhaps these statistics and our experiences in collecting them will lead to getting a more complete picture of our electronic resource use in the future.

Getting the Best Out of It! Usage Analysis from the Publishing Perspective

by **Marthyn G.M. Borghuis** (ScienceDirect, Senior Manager, Usage Research, Amsterdam, The Netherlands) <M.Borghuis@elsevier.com>

What Factors Drive Usage?

The conversion from print to electronic publishing is proceeding rapidly as is the retrospective digitisation of the archive of print articles. Availability of full text articles in electronic format does have a substantial impact on the volume of usage data generated, this is obvious. It seems, however, that for publishers the three crucial factors influencing usage levels are:

1. The number of universities, corporations, hospitals, and research organizations worldwide that have access to the online journal collection,
2. The number of users, within these institutes that have access to the collection, and finally,
3. The increase in the number of readings (and thus usage) by these users (usage intensity).

The Global Trend

Knowing these main usage factors, what then is the global usage trend across all publishers offering primary content online? Where is the sky and does it have limits? It is an opportunity to present to the readers of *Against the Grain*, a straightforward analysis of online journal usage developments worldwide to date and in the years to come. For that exercise the usage data of ScienceDirect have been used covering the period Jan 2001 to Sep 2003.

Analysing Elsevier's ScienceDirect data, the following assumptions can be made: Extrapolating from ScienceDirect usage and the size of its journal collection relative to the total scientific journal literature available online, it is likely that across all primary publishers at least 90 million requests for full text articles were made on a global scale in September 2003. In addition it is essential to know, that full-text usage is increasing by an average 15% in peak months (September to November & February through April) and about 8% in all other months. As a consequence, every year so far, usage has more than doubled and there is no sign whatsoever that this will decline in the next two to three years! It is thus expected that total full text article usage achieved in 12 months, across all libraries and publishers worldwide, will pass the **1 billion landmark** in early 2004!

Looking at this trend it is very clear, that the transfer from library print collections to electronic full text collections has resulted in a tremendous increase in accessibility and thus usage of full text scholarly and scientific articles.

In itself this is a true success story!

What is Contributed by each of the Usage Drivers?

a. *Number of Institutes and Users having access:*

A King & Tenopir (2002) analysis resulted in a count of 6.8 million scientists in the USA "engaged in research and development and teaching, who were likely to be readers of scientific publications."¹ Numbers are based both on professional scientists in academia, corporations, government, and research organizations. Knowing these US numbers, how large will the potential reader base be on a global level? There is evidence that the US portion is around 30% and thus that the worldwide number of scientists, R&D staff, and educators is approximately 20 million. After starting in 1999, the present size of the ScienceDirect customer base is 70% of the total estimated number of research institutes (i.e., institutes that maintain a library facility, which serves users, who read from scientific and scholarly

journals). As a library has a collection from various publishers, other primary publishers will also have reached those institutes. There might be a slight variance in the total numbers for each publisher, because publishers with a

small and narrow collection will have a lower number of potential customers.

b. *Usage-intensity:*

In their earlier study King and Tenopir found that readings in the non-academic sector amount to 102 per researcher per annum, and in academia to 188.² However, the authors report also that the number of scientists residing outside academia is a lot higher than that within academia. It is thus necessary to set the average readings at a lower level and estimate 135 readings per annum across all readers. The authors recently updated their findings and report a continuing increase in average annual readings per academic reader, up to 216 for 2003.³

Adapting this number again for scientists outside academia reading less, the average number of annual readings across all US scientists should be set at 153. Finally, the authors take another factor into account: about 60% of all readings are still from print, which leaves 40% from electronic sources. The equation is complete when estimating that about 70% of all institutes do have access to online articles. This means that total annual US readings from electronic sources can be found by the following equation: $153 * 6.8M * 0.40 * 0.7 = 290M$. This result is pretty close to the expected US 2003

continued on page 36