2003

The Beginning of Value Assessment: Usage Information in the E-Journal Age

John Sack
Stanford University, sack@stanford.edu

Follow this and additional works at: http://docs.lib.purdue.edu/atg

Part of the Library and Information Science Commons

Recommended Citation
DOI: http://dx.doi.org/10.7771/2380-176X.4203

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.
The Beginning of Value Assessment: Usage Information in the E-Journal Age
by John Sack (Director, HighWire Press, Stanford University; Phone: 650-723-0192) <sack@stanford.edu>

Right from the start of HighWire Press with the publication of the Journal of Biological Chemistry online in 1995, there was a strong interest in learning what we could from usage statistics, or “server logs.” We knew we should measure just about anything that moved or clicked, because the early days of e-journals were like the early days of space exploration: we didn’t know what observations would be useful because we didn’t know what new phenomena in user behavior would occur. With over 350 online journals, from over 150 different societies and publishers, about all the HighWire-hosted publishers could agree on was that we needed to measure “hits.”

At first, each journal editor wanted to watch the hits climb - and they did climb: a number of journals in the 1990’s saw usage increase by factors anywhere from 2 to 10 from one year to the next as some whole user communities rapidly shifted from print to online use. But soon after, editors realized that they had data on usage patterns that they couldn’t really obtain from counting print copies mailed, or even by reader surveys: what are readers reading?

This began our understanding that electronic journal usage information could be “mined” for more than more hits: it could tell those who have intellectual responsibility for collecting articles into journals - editors - and those with responsibility for collecting journals to support academic programs - librarians – what role a journal had in the intellectual life of a discipline or institutional community.

We also realized that the initial introduction of a new technology such as e-journals was the best and perhaps only time to instrument the shift in literature-research pro-
We began our analysis by asking librarians what type of information would be useful and/or interesting — the type of information that might provide its own insights! This was all being discussed before online publishers were regularly delivering usage information, before IFLC developed general standards, and years before COUNTER began to codify uniform practice. At the same time, we asked journal editors and publishers about their interests in measurement. We thought that public service librarians and journal editors might have similar interests and that selectors and publishers might have similar interests because each of these pairs serves a somewhat similar function but at different "ends of the pipeline." And in all cases we felt that the principle of transparency should apply: information needed by either community should in principle be accessible to the other.

Were did this lead us? Of course, we saw the need for counts of various types of actions such as searches and article displays, just as all

Figure 1: Six Core Information Practices

Because HighWire is located in the Stanford University Libraries, with a management team that has been working with libraries since the development of BALLOTS and RLG, we were naturally interested in how usage information might assist libraries in making selection and de-selection decisions — perhaps re-shaping the economics of journal publishing — as well as how the information might help librarians provide better public service. It seemed to us that the usage information contained in e-journal Web-server logs could provide a new source of information to librarians serving a more and more "remote" online community. That community could increasingly take advantage of library services without stepping into the physical library, and whose information needs could less and less be understood through personal interaction, much less through circulation and re-shelving statistics. This was especially true in the science-research community where HighWire got its start.

Number original content
Organize content
Circulate and exchange knowledge
Regular monitor and review
Directed research for retrieval
Intensive study reading

1) We found that there was an interest in understanding intellectually what people were reading: which articles were most often read each month and which sections or topics of articles were most often read each month. This kind of information tells an editor or a librarian what is important in an academic or research program. (Of course, the statistics do not track individuals' use, only aggregates.)
2) We found an interest in assessing reading for current awareness vs. archival functions. How old is the content people are reading? Is most of the reading material pulled from the current issues and volumes, or from older material?
3) How are researchers finding what they read? Are they locating articles by browsing through issues, or are they locating articles by known-item searching by author/title, or by subject searching? We believe, from our usage statistics, that they are predominantly searching, by about 2:1.
4) How broadly distributed is usage across a patron population, and across a database of articles? Are there many "light" users, or just a few heavy users of a journal? Are many articles being read across many issues, or is reading primarily in the current issue, or in just a few articles? The "unique events" counts in the HighWire usage statistics tell the story here.
5) How does academic-period usage compare across time? We found there was interest in comparing academic-term (or a range of months') usage in one year vs. another year, rather than comparing two subscription-years' usage. HighWire's "custom multi-month reporting" enables this function.

Of course, there were some "interesting" ideas that we couldn't support, such as reporting on IP ranges within a subscription. But maybe someday!

We heard from readers that they were interested in knowing what others were reading, specifically in the categories of their studies. This gave rise to the "Most-frequently Read" reports, which some of the HighWire-hosted publishers post on their Websites for readers to peruse, for example see Gut Online's most frequently read articles here: http://gut.bmjournals.com/reports/mfr1.dtl

And, of course, we heard from authors that they wanted to know how often their papers were read and cited. This is a feature of the New England Journal of Medicine Website. We even heard from a tenure-review committee asking for data about how often the papers of a particular researcher were read!

More recently, we began work on a tool to help libraries study their patrons' collective use of journals to which the library does not subscribe, and, reciprocally, a tool for publishers to identify non-subscribing institutions where there is high use of a journal. When we reviewed the idea for such a report with librarians, there was good agreement that this kind of information would be useful as long as the same information was available to both publisher and librarian. These new reports are just about ready for release to publishers and librarians. One might perhaps wonder where such use would be coming from if there were a lot of it, but the library didn't hold a subscription. Because there is so much free full-text content available from HighWire-hosted publishers - there are almost 620,000 free full-text articles available from 165 journals participating in the "Free Back Issues" program, making the HighWire archive the largest free full-text science article repository that we know of - there is a good opportunity for librarians and publishers to judge demand for a journal's content without ever having to implement and promote trial subscription programs.

Since the HighWire usage reports provide continued on page 40

<http://www.against-the-grain.com>
understand how valuable the content is in general to an academic community, because the use is dispersed through multiple "channels." There is no place or report that pulls all the use together of the intellectual content in a title.

As librarians try to decide whether an "aged" version of a journal in Ovid (a version whose newest issue is three issues back from current), or an incomplete version of a journal at Ingenta (a version that does not include non-print material), are sufficient to support a research and instruction program relative to the complete "journal of record" available at a publisher's own site, we may see that usage information that is insensitive to the age of content being accessed will not suffice.

Librarians may need better or different statistics to calculate the increasingly popular "cost per use" metric. Such a metric gives librarians — and publishers — a tool for validating that value is commensurate with cost. A good set of metrics in this regard should help with librarian/publisher negotiations, especially in complex multi-institution cases.

Ultimately, science and publishing will need an additional metric going beyond the "impact factor" as a way of judging the importance of a journal. While no one seems to have come up with a replacement for impact factor - it is tempting to say that like democracy it is the worst possible approach except for all the others - with the availability of much more fine-grained and pervasive metrics on usage, it should be possible to build alternatives that might, for example, more accurately indicate the value of the clinical journals whose articles are heavily used by practitioners, but not widely cited by researchers. Impact factor emphasizes one aspect of value over other aspects, and some of the others may now be visible and useful in usage reporting. For example, in the Web world where the impact of a site is measured by its raw traffic, a more useful statistic could be the total citations to a journal (i.e., the impact factor without the division by the number of articles in a journal), because that is the best predictor of the total amount of visitors a journal site will get; measured this way, the Journal of Biological Chemistry is the "highest impact" STM journal, because it is the most-frequently cited. And, it might be possible for librarians to calculate the impact factor of a journal relative to their own academic programs.

Librarians may want to find a way to use usage information to negotiate prices for products when portions of a product — such as a database containing journal articles — contain content that has already been purchased through another channel. This would require fine-grained analysis of usage of portions of a database.

Librarians may want to assess the amount of use that publisher archives of 20, 50 or 100-year-old articles have, so they know whether a continuing subscription to such an online archive is worthwhile.

Online journals provide services in addition to content. HighWire-hosted journals provide not only the usual email tables of contents, but also personalized (keyword and subject-specific) alerts when new content is published. Part of the value of a journal is in these services. Online journals are beginning to provide e-journal content in PDAs and via RSS (Really Simple Syndication) protocols to campus communities. Should the use of these services by a campus community be tracked?

Authors may want to know who cites their work — this information is available at many HighWire-hosted journals through a partnership with ISI on its Web of Knowledge product line — and even what institutions most often read their work and incorporate it into course packs. Certainly as publishers increase the flow of authors to retain copyright, we will see the authors' interests in the usage of their published increase.

Usage statistics are becoming an important way for a library, and for a publisher, to assess whether value matches cost. Perhaps as these assessments evolve we'll see the statistics evolve as well, beyond "usage" in the traditional sense. For example, under the new American Physical Society "blended pricing" model, an institution's frequency-of-authorship is taken into account in establishing the subscription price for a journal. Librarians may even want to identify authors to suggest alternative publications to faculty who publish articles in expensive titles. As additional delivery platforms become available to authors — such as institutional self-archiving programs starting at a number of institutions — these calculations of a journal's value will become more complicated.

Clearly, significant challenges and "interesting" times are ahead for us! Usage, content and value continue to be a dynamic equation depending on whose usage, whose content and whose value is at issue.