December 2003

Making Haste Slowly: E-metrics -- Where Are We Now and What Can We Expect?

Bob Molyneux

U.S. National Commission on Libraries and Information Science, bmolyneux@nclis.gov

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

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

Recommended Citation

Molyneux, Bob (2003) "Making Haste Slowly: E-metrics -- Where Are We Now and What Can We Expect?," *Against the Grain*: Vol. 15: Iss. 6, Article 6.

DOI: <https://doi.org/10.7771/2380-176X.4197>

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.

Making Haste Slowly: E-metrics — Where Are We Now and What Can We Expect?

by **Bob Molyneux** (Director, Statistics and Surveys, US National Commission on Libraries and Information Science; Phone: 202-606-9181) <bmolyneux@NCLIS.gov>

Using digital materials as a method to disseminate information is an exciting development for information professionals and academic researchers. The transformation brought about by the digitization of the human record is proving revolutionary since it makes the transmission of information so much easier than earlier formats.

Part of this revolution is trying to get a handle on what is going on for purposes of planning, budgeting, and collection development. In fact, we have many people chasing fragmentary data, looking for insights into what is going on while often faced with budget decisions that have to be made by next week.

We largely have anecdotal evidence about the general situation but there is a budding community of people focused on systematically measuring the use of materials in digital formats. Anyone who wants answers to questions right now is going to be frustrated because measuring the impact of these materials has proven to be a difficult problem. Better data is coming but even though the data will be better, it will never be perfect. There are sure to be un-

available or unreliable data for some time to come that will not fully alleviate these areas of frustration.

COUNTER

I have argued in *ATG* before that **COUNTER** is the best shot we have for getting better, comparable, generally available, and believable data than any other initiative. The first **Code of Practice** is sound, systematic, and reflects serious attention to the problem. It is the *de facto* standard and once the XML and auditing infrastructures are in place, then we will begin to have basic information of use to both the practitioner and the researcher. We will be seeing **COUNTER** compliant data soon but XML formats, as I understand it, are not obligatory before 2005. Expect to see some of these data before then, and see the **COUNTER** Web site for some example reports <http://www.projectcounter.org/>.¹

XML formatted samples also exist. **COUNTER** is an organization where a process for development and improvement of electronic usage data will continue.

COUNTER is now soliciting memberships to fund its work and its work on data exchanges standards deserves support from the library community.² **COUNTER's** existence, though, will not end the quest for data for a set of reasons that are endemic to the problem at hand.

The first and easiest reason to understand is that there is a hard data problem since just about everything that makes data collection difficult exists: the processes being measured are complex, changing, and new. Data collection is easiest when you are collecting data on simple, stable, and old processes. So, even with the best intentions, there are technical difficulties that will require work to solve. Some, I fear, may prove intractable.

Cooperation and Contention

The second reason that **COUNTER** will not end the quest for data has to do with the different objectives the various parties bring to their quest for data. Data are not collected in a vacuum but to meet the objectives of those collecting them. These differing objectives create centrifugal forces that will bid to pull the **COUNTER** effort apart.

Librarians will not want the same things as publishers, so there will be areas where there will be agreement and other areas where there will be profound disagreement and we will all just have to get along if it is to succeed. For instance, librarians look at these collections of serials titles that they are paying for and look

for means to cut some of them or to rationalize the selection process so they can plan for the future and can tailor their collections to their users. From the librarian's perspective why have access to a given title from five different places? Why not get access to that title from the cheapest place? And with comparable use statistics from **COUNTER**, some titles could well be canceled.

Publishers and aggregators, of course, do not like the idea of cancellation so they have an incentive to prevent publication of some types of data or obscure others. We all may be better off in the long run if libraries can cancel some titles but each of us won't necessarily be better off tomorrow. Hence, it is arguably rational for those firms supplying data on use of electronic materials to supply non-comparable data because it makes comparisons that might lead to cancellations difficult. Similarly, it also seems rational to restrict sharing of contract terms and sharing of data between institutions: why inform the library that paid the most for a title that everyone else paid less? Why tell everyone which library got the best contract and have that become the floor for all future contract negotiations?

Neither community is monolithic, however. The library community has parties within it with different objectives that could also disagree with each other about **COUNTER**. I have had librarians say to me that they do not want data published because it would make them "look bad" so some may be ambivalent about various data items that are discussed in **COUNTER**.

Of course, those in the publishing and aggregator sector also have crosswinds blowing in their communities. Those who have read the **Morgan Stanley** report on **Reed Elsevier** know that clear-eyed appraisal of the situation libraries are in and how it can be exploited. It is tempting to think of this report as the blueprint for all publishers and aggregators. However, some publishers or aggregators are stronger than others and some are under more pressure than others to increase revenues, gain market share, or whatever. Some may take a long view while others a shorter one. These differing situations will lead to different attitudes about what should be reported.

Some publishers and aggregators must also be concerned about overreaching because if they go too far in increasing costs, the politicians get involved and then all bets—and business plans—are off. These facts in their environment will mean that they will disagree from time to time with each other both in which data to supply and in which data they can afford the programming to supply.

Now, none of this is either bad or good as long as everyone understands the different objectives. Collecting these data are difficult by

continued on page 20

Electronic Use Statistics: from page 1

Two publishers are represented among our authors. **Marthyn Borghuis** of **Elsevier ScienceDirect** describes the factors that influence usage of online journals and explains what publishers do with the usage data they collect. **John Sack**, of **Highwire Press**, explains the usage studies that Highwire has completed and the findings that their extensive analysis show about online journal use.

Both standards-based projects for collection and reporting of usage statistics are represented in the issue. **Denise Davis** brings us up to speed on the **NISO Committee AY** and the **Z39.7-2002** standard in her article, while **Peter Shepherd** describes the progress that **Project COUNTER** has made to date and its plans for the future.

I hope that all our readers enjoy this issue of *ATG* and find each article as thoroughly interesting and thought provoking as I did! And remember, we can look at statistics two ways: as **Paul Fisher** believes, "A passion for statistics is the earmark of a literate people," or as former major league pitcher **Jim Bouton** believes, "Statistics are about as interesting as first base coaches." 🐘

Making Haste Slowly: from page 18

their nature and not the least of that is a result of the political fact that many parties are coming to the table with different objectives. Fine, so we will start with a set of least common denominators. Let's work to make **COUNTER** the best least common denominator we can and build on that foundation based on our congruent interests.

Sometimes these incentives will lead us to cooperation and sometimes to contention but, in any standards body, this kind of contention occurs, particularly as we have seen in Internet standards bodies where commercial and not-for-profit interests have sometimes clashed sharply. The goal is better data and if we work toward that end in a disciplined fashion, we will have a centralizing force to overcome the centrifugal forces these disparate objectives will generate.

OTOH, Let's Keep the Powder Dry

There are generally two kinds of uses for data: decision support or research into relationships leading to understanding and prediction. In a practical field like the library field, there should be overlap but the record in the field on analyzing data of any kind is dreadful. Library science is not in any real sense a science.

That said, it is impressive what is going on in the measurement of use of electronic materials and how people interact with them. In fact, there is probably a sufficient body of material for a useful bibliographic essay of the scope of those found in the *Annual Review of Information Science and Technology*. I think we actually know a fair amount if we had everything collected in one spot... here is a job for evidence-based librarianship. I have a sense we know a lot more than we think we know if it were all collected and assessed.

While there are articles based on data gathered in the libraries from direct measurement, most articles one sees use vendor-supplied data. The kind of data gathering outside vendor-derived data is taking place in institutions where there is technical expertise. OhioLINK <http://www.ohiolink.edu/>,³ and the **University of Pennsylvania**⁴ are two I know about but I am sure there are others. In these institutions, the actual streams of data are being analyzed and these institutions are not depending on vendors for all their data. I think these kinds of independent initiatives are a vital development.

Here is a place where individual use of materials may finally give us a means to measure the effect of libraries on people directly while respecting the privacy of the subjects. Here, too, data that will not be available from the **COUNTER** process for whatever reason may be investigated. Maybe you won't be able to get this or that important number from the vendors but maybe you can get ideas of what those values would be from other sources or perhaps these institutions will make software available to do the work locally.

What next?

There is an interesting mix of factors in providing digital information that will affect the future in the near term but this is the beginning

of a revolution that will take centuries to work out. That fact means that there is bound to be more change and more difficulties with data in the future. Data for the librarian making selection decisions will start to get better quickly, though.

I blanch to say this, but on occasion the library world has relied on the kindness of strangers to collect its data and the experience has rarely been a happy one; an independent effort should be maintained. We do not want a monoculture of data on the use of electronic resources so that kind of effort is important to keep going. The independent effort should continue for another reason: **COUNTER** may fail and independent efforts outside vendor-supplied data may be what we are left with.

I think we in the library community should do everything we can to help make **COUNTER** a success and being aware of the many varying objectives of the parties involved is a way to help do that. Another way to keep the centrifugal forces in check is a countervailing force: commitment to the objective of collecting data we all trust so that we can make this new enterprise we are all engaged in work. We do not want to kill this baby in the crib.

Column Editor's Note: Thanks to **Phil Davis**, **David Goodman**, and **Chuck Hamaker** for their observations and comments. Errors are mine, of course. These opinions are my own and not those of the **National Commission on Libraries and Information Science**. — **RM**

References

1. **COUNTER's** Web page is at: <http://www.projectcounter.org/>. Examples of usage reports are at: http://www.projectcounter.org/code_practice.html#section4.
2. Information on **COUNTER** membership is at: <http://www.projectcounter.org/member.html>.
3. OhioLINK's Web page is at: <http://www.ohiolink.edu/>.
4. Zucca, Joe, "Find a Way or Make One: Strategies for Measuring and Analyzing the Use of Digital Information," PowerPoint slides from a talk at the ALA Annual Meeting in Toronto, June 20, 2003, <http://www.arl.org/stats/newmeas/emetrics/arljune2003.ppt> (14MB).

Bradford's Distribution, the 80/20 Rule, and Patterns of Full-Text Database Use

by **Steve Black** (Reference, Instruction, and Serials Librarian, Neil Hellman Library, The College of Saint Rose; Phone: 518-458-5494) <blacks@mail.strose.edu> and **Amy Sisson** (Neil Hellman Library, The College of Saint Rose) <sissona@strose.edu>

A well-established principle of bibliometrics is that a relatively small number of journals get the majority of use. A simple mathematical formulation of this principle is the 80/20 rule, which states that approximately 80% of journal use is concentrated in 20% of the journal titles (Trueswell 1969). The 80/20 rule may be applied to online journal use as well as print use; a recent study of consortial use of online journals in the **Academic Press IDEAL** aggregate package found that 24.3% of titles satisfied 80% of downloads (Davis 2002).

A more elaborate mathematical model for the concentration of journal use in relatively few titles is **Bradford's** distribution, stated as $1:n:n^2:n^3:n^4\dots$, which **Bradford** originally derived by examining the use of print journals within the specific disciplines of applied geophysics and lubrication (Bradford 1953, p.154). This distribution of use, also known as **Bradford's** law of scatter, is recognized as fundamental

to bibliometrics and information science, and has received significant attention since its publication (Lockett 1989). **Nisonger** (1998) provides an overview of **Bradford** distribution as it fits into the bibliometrics of journal use, and **Oluic-Vukovic** (1997) has described the development of mathematical models based on **Bradford's** law of scatter and related distributions.

To determine **Bradford's** distribution for a set of journals, the journals are ranked in use order, with the journal receiving the highest number of uses ranked first. The **Bradford** multiplier, n , is equal to how many of the next highest-ranked journals it takes to achieve the same number of uses as the highest-ranked journal. Thus, if the first journal receives 1,000 uses, and the next three ranked journals together receive approximately 1,000 uses, the **Bradford** multiplier n is 3. **Bradford's** formula then predicts that it will take the next n^2 journals, or 9 journals in this case, to receive the next 1,000 uses; the next n^3 , or 27 journals, to receive the next 1,000 uses, and the next n^4 , or 81 journals, to receive the next 1,000 uses.

During **Bradford's** day, it was pro-

continued on page 22