February 2003

Devil's Advocate -- COUNTER, XML, and Online Serials Measurement

Robert 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>

Recommended Citation

Molyneux, Robert (2003) "Devil's Advocate -- COUNTER, XML, and Online Serials Measurement," *Against the Grain*: Vol. 15: Iss. 1, Article 39.

DOI: <https://doi.org/10.7771/2380-176X.4273>

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.

International Dateline from page 91

The other important item on the calendar is the so-called "Amsterdam" meeting of STM in the late spring. In 2003 it is back in Amsterdam.

In London in December the **International STM Association** hold two seminars (information-sharing) at the end of the Online week. The first of the seminars, on the Thursday, is mainly for those involved in the internal processes of the publishing houses and is entitled *Strategies of Journal Production in the Electronic Environment*. Librarians will be pleased to learn that publishers really are trying to get digital versions of journals articles online, in correct form and before the print version is available. It is a very difficult task, however.

The seminar on the Friday is the *Innovations Seminar*. This year there were over 90 present, mainly publishers, and the theme was web services—the third wave of online publishing. What is this third wave? The definition provided is that "publishers begin to develop innovative web services that add significant value to their content" and goes on to predict that "with the advent of the third wave, the journal itself may begin to be transformed from the model that has

I, User from page 87

can help make sense of separate conversations with Collection Development, Acquisitions, Serials, Cataloging, Database Maintenance, Systems, Preservation, and consortial partners.

5. The challenge for vendors/libraries is to learn how to shoot the curl of the information wave rather than be pounded into the ocean floor. (The challenge for writers is to avoid sentences like that one, but it's 1:10 AM, and the deadline was at midnight.) Both customers and those who serve customers (and we are all both!) must work at the coordination, communication, and information sharing.
6. Consultants!

As the customer experience becomes more complex and frustrating, there is some risk for all providers of goods and services, including libraries. If patrons or customers receive unhelpful or partial responses or have to extract needed information piece by piece, they'll eventually give up, and will stop using that vendor or library. They'll switch to a competitor (**Google!**), or stop looking, or revert to the time-honored expedient of watching TV all the time. (Assuming, of course, that customer service can get their digital cable working.) You don't want that on your conscience now, do you?

But there is more good news here than bad. Now that technology, customization, and strategic partnerships have become the very air we breathe, competition can and will shift back toward service, toward helping customers and patrons get what they need from our products and services. The winners will be not only customers, but those organizations that can understand and simplify the customer experience.

served the scientific community for hundreds of years." In library circles, a prediction like this would refer to so-called "alternative publishing," the cluster of ideas and initiatives based around SPARC. What does it mean for STM publishers?

The keynote speaker was **Patricia Seybold**, the eponymous leader of the group of that name. Her line is making it easy for customers to do business with you. Her thesis is developed at <http://www.pspgroup.com>. What publishers offer should encourage rather than deter publishers—not rocket science but, as I am sure librarians will feel, something that needs to be said and at length. The other presentations, from publishers, developed the theme and personalization of products and services and forging closer relationships with the author community. For those interested in delving into such matters, the presentations themselves may be obtainable through <http://www.stm-assoc.org/infosharing/presentations.php>. Among the interesting offerings were, as usual, innovations from the **Institute of Physics Publishing**, who provide an annual update on what they are doing at **Charleston**. They propose customer focused services for authors and referees, recommendation services like **Amazon** and the creation of a smoother interface to encourage users to mine the content on offer. The speaker from **IOPP** pointed out that some of their innovations had not worked. He instanced, as not wanted by the community, special interfaces and dedicated "filing cabinets." The **Elsevier** presentation concentrated on their investigation into

what journal authors want and how they can satisfy these wants. Central to developments is the quest to provide authors with a 360 degree view of the publishing process so that the author can find out where his or her paper is in the process with ease and 24x7. "Traditional" publishing works because authors (not to mention editors and referees) go to traditional publishers: a lot of work is going on sorting out those practices that exasperate authors—and editors. It is clear that **HighWire Press** is, at last, beginning to view itself as more than just a facilitator if a little less than a publisher. As many readers will already know there is a lot of work going on in Stanford to provide offerings on behalf of all or most of the publisher clients. One perception interesting to me that came from the presentation by **Richard Newman** is that users need to have lots of options, but be given simple routes. Like Swiss Army knives, all tools should be available but not all open at the same time.

This is a quick selection of innovations. The discussion brought up, mainly tangentially, some concerns and some fundamental queries. Are we talking about content or services? What will happen to the journal itself in the "article economy?" Can publishers really handle the supplementary material in digital form, which is now beginning to stream in and which takes up a lot of space and handling? An old stager from the floor pointed out that librarians do not have money to pay for all the potential services on offer. I imagine that many readers will concur.

Devil's Advocate — COUNTER, XML, and Online Serials Measurement

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

The purpose of this article is to elaborate on the major points of my talk at the last **Charleston Conference** and to discuss recent developments.

BACKGROUND

There was much discussion at the Conference about the measurement of the uses of online journals and databases. All agreed on the importance of this question and the inadequacy of the current reporting of data on the use of these expensive materials. **Ann Okerson** ably summarized these problems so I defer to her discussion of the many issues involved.

Given that there is a demand for the collection and distribution of "good" data on the use of online materials, the question now is: how to do it?

I have concluded that the true cost of that unpleasantness in the Garden of Eden was bad data. Before Adam and Eve messed things up, data were good and readily available. Now, we have to work to get good data—by the sweat of our brows. It is hard and expensive work.

THREE STORIES

Probably the prevalent view in the field is that good data come from sufficiently clear definitions. Let me share with you three cautionary tales from work I did with the **Association of Research Libraries (ARL)** in the 80s.

1. Expenditures for automation

At the time, the large academic libraries that were members of **ARL** were committing increasing resources to expenditures for automation. It was important to find out how much was being spent but there was no category in the **ARL** data to collect this number. **ARL** had not collected data on automation expenditures because it was a new and rapidly developing area. Like the matter of the use of online journals and databases, it was an expensive and fast-growing target.

Two committees of **ARL** each independently decided to find out the answer to the question of how much was being spent by the **ARL** li-

continued on page 93

libraries and each committee independently designed forms and mailed them to the member libraries. Typically the member libraries have one person to answer ARL statistical surveys and they dutifully responded. I wondered if the two sets of numbers were statistically related and analyzed them and found, basically, a weak if any statistical relationship between the two sets of numbers. My papers are still in boxes from the move so I cannot cite the exact numbers but that was the result. In other words, the same organization surveyed the same libraries for information on the same phenomenon and there was almost no statistical relationship between the results.

2. Ph.D. fields

ARL asks its members for the number of Ph.D. fields each academic member's parent institution has. These fields were defined by a form from the Department of Education at the time. In addition, two other widely available reference books also reported on which fields—as defined in the same Department of Education form—offered doctorates. I counted the number of fields in those two publications for ARL university members and compared them to the published figures from the ARL data. Was there a relationship with the three sets of figures reporting the number of Ph.D. fields? No.

Same institutions—usually from the same institutional research office—reported data to three different places purportedly based on the same Department of Education definitions and there was little statistical relationship between the three sets of numbers.

3. Administrative staff

ARL does occasional "Supplemental" surveys. There are several purposes for these surveys such as testing new variables for possible inclusion into the annual survey or to get a handle on a question that is not likely to require an ongoing effort. The question arose: what percent of ARL staff are administrative? A large number of libraries reported.

At a subsequent meeting, one director had strong words to say about the statistics reported by another. I am making numbers up but the director at one library had reported something like 30 people involved in administration and the other had reported three. The libraries were similar in size and the one who reported 30 was not happy with the one who reported three. After all, which library looks more efficient? Here, I infer, they each interpreted the question differently and I bet if one were to have talked to each of them separately at the time, you would have heard two different—but I bet reasonable—interpretations of the instructions on the survey form.

TAKING A STEP BACK.

I wish to make two general points about these stories.

First, from these events and others, you can see that getting good data takes time and takes more than good definitions. In coding data forms responding to surveys with those definitions, those filling out those forms typically note where the reported number differs from the definitions and for a variety of reasons. These notes the compilers dutifully report in the published compilations and these footnotes then disappear from the electronic versions of the data.

In any case, sometimes it is the working out of differences as the process illustrated in the story about the number of administrative staff members at those two libraries. You can often observe this process by following new variables as they are added to reported data through time. Over time, the variation between the threes and thirties will decline as individuals working on data discuss with colleagues at other institutions details of how the data are compiled and reported. Other kinds of work are involved as people tinker, talk to each other, and notice oddities.

So, time is a part of the equation when the data are collected regularly and close definitions as we find in the other two events do not cause reported data to be reasonable.

Second, given the length of time it takes to create the infrastructure necessary to generate data that are reasonable, how do people in other fields get such data quickly? They sample their populations. Sampling was developed as a means of getting data when getting population data are expensive, impossible, or when it is needed quickly.

Librarians seem to sample seldom. There are examples, of course, but it seems that usually when we librarians set out to collect data, we try to collect everything we can think of thereby, I infer, in an attempt to collect what we hope will be perfect data. We try to hit a home run every time at bat. Alas, data are a good servant but a terrible master and people respond to long forms by answering fewer questions. The reaction to long forms causes what is called "responder burden." The highest response rates I have ever seen are to postcard surveys (you mail the responses back on a postcard) with a few questions. People answer those quickly.

We have, then, two major kinds of data collection: samples done where collecting data from all members of a population is difficult and systematic, and long term collections of data such as the ARL or Federal State Cooperative Statistics (FSCS) data. Time series have proven to be the most useful but they are also the slowest and most expensive. And often they are not appropriate to the matter at hand. Sometimes quick responses now from a sample that are representative of the population are of more use than perfect data next year.

MORE ON THE PROCESS OF GETTING GOOD DATA

We also have two major uses of data, in addition to the two major kinds of data collection.

I have discussed before the "support" and "illumination" functions of data in other issues of *Against the Grain*. To summarize, uses of data for support are those uses for decision support such as justifying budgets. Most fields also have the illumination or research function where the "is" rather than the "should" is investigated. Bob Williams, at the University of South Carolina's School of Library and Information Science would add "problem centered" to this taxonomy because he sees in this term a kind of thing not covered in the other two terms. Problem centered would be a research function where you might ask: what are the precursors to the development of public libraries and then look for or gather data to help answer this question.⁴ I include this kind of research in the illumination function here for simplicity.

In any case, the "illumination" function is vital to the production of good data because people who do this function investigate and improve data. Good data is the result of a process, not a one-time event and lacking this illumination function means that, in effect, library data collection—as a systematic discipline—is rather like the sound of one hand clapping. We do collect some pretty good data—witness the ARL data or the FSCS data to name but two, but they are rarely analyzed. As mentioned, though, we are not so good at things which arise quickly and expensively—witness expenditures for automation and the thing to which I now turn: online materials use.

ONLINE MATERIALS USE IN ITS DATA CONTEXT

The importance of the question has led to several efforts to improve data collection on the use of these materials. Peter Shepherd spoke at the conference about Project COUNTER², one such effort and one with substantial vendor and publisher backing. Of course, the International Coalition of Library Consortia ICOLC³ and ARL⁴ are also involved in their own efforts. NISO's Z39.7⁵ draft standard does not deal with this question directly but does have variables that touch on it. It is hard to know how all these relate to each other but, in the computer and networking world, there is a saying that God must have loved standards because he created so many of them. I'm with Rodney King on this one: "Can't we all just get along?"

I have been skeptical of the traditional approach of committees drawing up carefully wordsmithed definitions and publishing them, then leaving the field to make sense of the result, for reasons adduced above. The latest Z39.7 standard and Project COUNTER, I think, present better approaches. Z39.7 uses existing work of other entities and Project COUNTER does three things I think are more in keeping with how data are developed when you are about it for the long haul.

First, its new Release 1 of its "Code of Practice" starts small by picking a few variables which it hopes to survey well. Rather than being beguiled by the prospect of collecting everything one could think of, they try, rather, to do a few things well. They may not hit home runs but they will get people on base. The longstanding ARL series started with a few variables which have been added to slowly over the years so the model has good precedent. In time, COUNTER will add variables as bugs are worked out of the system. It initially concentrates on online serials and on databases and provides for a number of demographic variables which I think are going to be important.

Second, COUNTER says it will monitor how the data are collected and tweak definitions as necessary.

Third, it has solicited comments from the field. This is more of a bottom-up approach than the traditional top-down approach. Peter Shepherd tells me that the first draft of this release had more variables but the comments they received led them to cut down the number for the final version of this release.⁶

Now, I am skeptical about some aspects of COUNTER. For instance, I can't wait to see if vendors will share the information on their product's usage given the reports I hear that vendors in their contracts with libraries specify that those libraries are not to share data on the use of those vendor's products.

continued on page 94

Devil's Advocate from page 93

Also, **COUNTER's** plan to require that results reported by vendors will be audited to comply with the Code of Practices is fascinating. Library data as submitted are traditionally unaudited so we have no precedent but the corporations these vendors, typically, are used to having their financial figures audited so this requirement may not be as odd to them as it may seem in the library world.

Now, I know, that those who want perfect data today will be frustrated by the number of variables that **COUNTER** begins with relative to the hundreds that might be collected but the plan is realistic, systematic, and serious.

I think the concept is good but execution will be another question. For that, we wait. We wait through the first few months or years as bugs get worked out and bad comparisons are made and we get used to what we have and, drat, why did they collect this variable and not that one? That is the way these things go. And we wait as too few try to sample aspects of the problem and write the results up. For a field that organizes the recorded output of the species we rely on Homeric methods—word of mouth—to pass on our information.

Acceptance of **COUNTER**, ultimately will probably depend on libraries and acquisitions librarians who will have the option of requiring vendors to comply with the Code. However, truth to tell, it will probably take five years for all the bugs to work out to the point that the data from **COUNTER** make the most fastidious people happy. Some of us will be happy the first year, though. And the more who use the data, the better they will ultimately be.

While this approach will take time, to get enough samples to make sense of what is going on, though, would mean that the library field would have a critical mass of data analysts and it does not. This is the dilemma. What will happen is that it will take time for this all to work out and the demand is now. What will we do in the meantime?

A BROADER CONTEXT

Decision support is the point that people are focussing on now because prices are going up and budgets are going down and decisions about what to retain are being made without data or with spotty data. This is an important question to which I will return presently. There are,

though, two broader questions I wish to elaborate on because these broader questions affect the kinds of data we should collect.

Library data has historically dealt with plant, infrastructure, and what has come to be called "inputs." What we have not been able to do is to measure directly the effect of libraries on people. In the **E.G. Holley** festschrift, I proposed calling the minimum measurable level of this effect the "Holley"—think of it as the subatomic particle of thought.⁷ This number would be an interesting one to have and we can begin to approach it if we can measure which articles researchers download and which show up subsequently in cited research. Crude, to be sure, but a start to an ultimate question of the impact of libraries on the populations they serve.

The second vital question is the information policy question. As a free society, we rely on an informed citizenry. How does this citizenry stay informed? What is the balance of rights of citizens as opposed to those who own or control intellectual property? That answer is a partial function of how we get that information and of how much it costs. To know that answer, we need aggregate, national data, in addition to data generated from individual institutions. It is important for you as a librarian to get a handle on what these resources are costing you but it is vital for the nation to know how much is spent and the result as it is also for consortia and for state governments. These resources are costly and agencies with a fiduciary responsibility need to have a handle on the scope and results of the expenditures.

The results of these observations lead us into two directions that were my major points in my talk. One deals with the inclusion of demographic data in vendor reports with **COUNTER** (or any other such standard) variables and the second deals with using XML as a standard method for reporting and exchanging statistics on libraries.

DEMOGRAPHIC DATA

I am using the term "demographic" to refer to variables that characterize not the actual use of materials but rather those variables that allow analysis and aggregation beyond one institution. Let me elaborate.

Given that national data on online materials use are important, it will be necessary to be able to retrieve data on institutions by type (public, academic), by geographic area (country, state or province, electoral district), size (by budgets or collections) and other such variables. I was

concerned that these kinds of items were not to be included in the **COUNTER** Release 1 and, I hope, they will be included on subsequent **COUNTER** releases. Like most who will look at this effort, I have my candidates for variables to be added in Release 2 and, wisely, they did not listen to all such suggestions for Release 1.

I was also concerned to make sure that data could be aggregated by vendor, journal, and the like, and I am glad to see these kinds of demographic variables are included.

XML

A problem we have now in our data is that too often, to use them for anything, you have to learn something else complicated. To work with many data, you might be best off learning SAS. But what if you just want to know something simple? One way to increase the number of people analyzing data is to make using them easier.

With the vendors producing data in so many formats—this one a Web page, that one an email—what do you do if you want to merge them? **Chuck Hamaker** has a nifty Web page⁸ that shows a start but this does not solve the underlying problem and that is the requirement for a general, flexible, and adaptable data exchange format. If different vendors wish to submit reports in whatever format and the title is always tagged in XML tags as `<journaltitle>Against the Grain</journaltitle>` people can write programs to fetch the data allowing those who want to use the data to put them into Excel or whatever. Individuals will not have to know SAS to make sense of the data; they merely need a question, a source of data, and utilities to arrange the data in a format they prefer. I plan on working on this question. It is complex but the result, ultimately, is more readily usable data. More about this anon.

CONCLUSION

I think that the matter of getting usable and useful data on the use of online materials is on its way to being solved but it will take time. The question at hand now is: what can we do in the meantime?

The views expressed here are mine alone and not necessarily those of the National Commission on Libraries and Information Science. — RM

Endnotes

1. Robert V. Williams. Personal communication.
2. Project **COUNTER's** Web page is at: <http://www.projectcounter.org/>. Release 1's page is at: http://www.projectcounter.org/code_practice.html.
3. ICOLC's Web page is at: <http://www.library.yale.edu/consortia/>.
4. ARL's Web page on its Measures for Electronic Resources (E-Metrics) effort is at: <http://www.arl.org/stats/newmeas/emetrics/index.html>.
5. National Information Standards Organization (NISO). Standards Committee AY: http://www.niso.org/committees/committee_ay.html.
6. Peter Shepherd. Personal communication.
7. Robert E. Molyneux, "More Hortatory than Factual: Fremont Rider's Exponential Growth Hypothesis—and the Context of Exponentialism," *For the Good of the Order: Essays in Honor of Edward G. Holley*. Delmus E. Williams, et al. eds. (Greenwich, JAI Press, 1994), pp. 85-117.
8. <http://library.uncc.edu/tech/stats/e-resourceusage.htm>.

ADVERTISERS' INDEX

61 ABC CLIO	91 BROCKHAUS/GERMAN BOOKS	69 KLUWER ACADEMIC
48 ACS	83 THE CHARLESTON ADVISOR	75 MAJORS
49 ACS	10 THE CHARLESTON REPORT	59 MCFARLAND
2 ALFRED JAEGER	27 EASTERN BOOK COMPANY	96 MIDWEST LIBRARY SERVICE
41 AMBASSADOR BOOKS & MEDIA	23 EBSCO INFORMATION SERVICES	45 OXFORD UNIVERSITY PRESS
31 AM. INSTITUTE OF PHYSICS	13 EBSCO PUBLISHING	95 PROQUEST
35 ASME INTERNATIONAL	17 ELSEVIER	33 SERIALS SOLUTIONS
5 ATG	25 ELSEVIER	7 SPRINGER-VERLAG
18 BASCH SUBSCRIPTIONS, INC.	29 EMERY PRATT	15 SWETS BLACKWELL
11 BERNAN	37 GREENWOOD PUBLISHING GROUP	55 WILEY
3 BLACKWELL'S BOOK SERVICES	71 GREENWOOD PUBLISHING GROUP	19 YBP
9 BOOK HOUSE	21 IEEE	43 ZUBAL BOOKS

Ad Manager: Edna Laughrey Internet: elaughrey@aol.com
phone: 734-429-1029; fax: 734-429-1711