November 2002

Devil's Advocate -- Variations on a Theme: NISO Z39.7, Library Statistics Standard

Bob Molyneux

University of South Carolina, School of Library & Information Science, drdata@molyneux.com

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

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

Recommended Citation

Molyneux, Bob (2002) "Devil's Advocate -- Variations on a Theme: NISO Z39.7, Library Statistics Standard," *Against the Grain*: Vol. 14: Iss. 5, Article 28.

DOI: <https://doi.org/10.7771/2380-176X.3777>

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.

SHARE OR SUPPRESS?

by **Pamela M. Rose** (University at Buffalo)

Should unclassified research that might conceivably help bioterrorists be openly published? Some members of **Congress** are calling on journals, scientists and funding agencies to rethink open publication of potentially sensitive microbial work, an issue biologists have been discussing for months. Although scientists agree with the premise, they are understandably wary of suppressing data.

See — “A Call for Restraint on Biological Data” by **Jennifer Couzin**, in *Science*, Aug. 2, 2002, pp.749-51.

SOMETHING OLD, SOMETHING NEW

by **Pamela M. Rose** (University at Buffalo)

Although new radiocarbon dating suggests the parchment on which the Vinland Map is drawn was made between 1423 and 1445, chemists argue that the ink is modern, thus making it hard to explain how the map might have been faked. Found in a private library in 1957, it is purported to be the first chart of the New World.

See — “Old Map, New Ink?” in Random Samples section edited by **Erica Goldman**, in *Science*, Aug. 9, 2002, p.931.

PATIENT PRIVACY

by **Pamela M. Rose** (University at Buffalo)

In response to growing concerns about access to medical records, the **Bush Administration** announced the “privacy rule” giving patients more control over how their records are used by health care providers. The rule, scheduled to go into effect April 2003, was modified to appease researchers concerned that data would be stripped of so many pieces of information that it would be unusable.

See — “Researchers Welcome Revised Privacy Rules” by **Jocelyn Kaiser**, in *Science*, Aug. 16, 2002, p.1109.

EVERYTHING INTO THE E-POOL

by **Pamela M. Rose** (University at Buffalo)

A bold come-from-behind strategy intended to make the **American Geophysical Union (AGU)** the next wave in science publishing has backfired in crippling access to articles in their print journals. As AGU went all out to make their online papers publications of record, they abandoned sequential page numbers in the paper version, making their articles impossible to locate from a citation. They have since adopted a four-digit article number. However time will tell if they survive their full-body-immersion into e-publishing.

See — “Online Pioneer Winds Up Lost in Cyberspace” by **Rebecca Renner**, in *Science*, Aug. 30, 2002, p.1469.

YOUNG RESEARCH

by **Pamela M. Rose** (University at Buffalo)

If you're still an undergraduate and frustrated by too many hours in class on not enough practical experience, **NASA** and the **National Science Foundation** may have just the ticket: adventurous undergraduate research opportunities designed to encourage students to choose careers in science and engineering. Experience a trip on **NASA's** “vomit comet,” or pursue galactic questions at the **Cerro Tololo Inter-American Observatory** in La Serena, Chile. It's hard to know if such programs really work, however most feel they do reinforce career decisions to go into science.

See — “New Lure for Young Talent: Extreme Research” by **Robert F. Service**, in *Science*, Sep. 6, 2002, p.1633.

INTELL PROP FLEES THE TAXMAN FOR SUNNY BERMUDA

by **Bruce Strauch** (The Citadel)

Multinationals are parking trademarks and patents in Bermuda subsidiaries so royalties earned outside the U.S. can be parked there tax-free. The lucre sits until it's “tax efficient” to move it to the U.S.

The hitch is the price you charge the subsidiary to buy a big share in the intell property. You want it low to reduce capital gains taxes. You're buying an intangible and the price is subjective — but the purchaser is a subsidiary. It's not an arms-length transaction.

And the IRS is all over this with billions in tax claims.

See — **Glenn Simpson**, “A New Twist in Tax Avoidance: Firms Send Best Ideas Abroad,” *The Wall Street Journal*, June 24, 2002, p. A1.

BRAVE NEW WORLD REDUX

by **Bruce Strauch** (The Citadel)

MIT is building an ambitious superarchive called **DSpace** where research material from every prof will be collected and tagged with metadata codes. Profs keep copyright and select access levels for contributed items.

This is yet another university effort to break out of the cycle of paying for research and then paying journal publishers for access to it. Proponents claim it will “unbundle” publishing leaving peer review to the publishers and distribution to the universities.

The scientist boycott of publishers who refused free online access six months after pub date flopped. The pre-print archive effort got nowhere. And Elsevier predicts superarchives won't work either. It is confident of the need for publishers to efficiently organize the “bundle.”

Besides, some professors fear universities won't be able to resist charging faculty for access. And there's the curious quote from **Caltech's** provost that in the superarchive future, journal publishers would charge authors rather than subscribers for their services. Hmmm.

See — **Jeffrey Young**, “‘Superarchives’ Could Hold All Scholarly Output,” *The Chronicle of Higher Education*, July 5, 2002, p.A29-30.

Devil's Advocate — Variations on a Theme: NISO Z39.7, Library Statistics Standard

by **Bob Molyneux** (University of South Carolina, School of Library & Information Science)
<drdata@molyneux.com>

---

1. NISO

The **National Information Standards Organization (NISO)** is in the process of developing the new version of the “standard” for library statistics (**Z39.7**). **Katina** asked me to attend the session at the **American Library Association** convention in Atlanta on June 14, 2002, where the **NISO** committee

gave a status report on the standard, and to report on the session for *Against the Grain*. This article reports on that meeting and discusses an idea that resulted both from the meeting and several conversations I have had with **Katina** and others about measuring the use of electronic resources.

NISO was founded in 1939 and is “...a

non-profit association accredited by the **American National Standards Institute (ANSI)** [that] identifies, develops, maintains, and publishes technical standards to manage information in our changing and ever-more digital environment.”¹ The first **NISO** library statistics standard was issued in 1968 as

continued on page 86

Z39.7-1968. Then came **Z39.7-1983**, **Z39.7-1995**, and the current revision, **NISO Z39.7-2002**, which is now published on the Web and is in a trial phase through July 2003. **NISO's** Website is at: <http://www.niso.org/>. The **Z39.7** data dictionary Web page is at: <http://www.niso.org/emetrics/>.

This revision differs from the other versions in two major ways. One already mentioned is publication on the Web, which is clearly more useful than a print version. Second, it includes "standards" on electronic resources.

The meeting was well attended; because there were not enough chairs, people sat on the floor and stood in the hall outside the meeting room. As one who has worked with library statistics for a long time, it was heartwarming to see that many people interested in hearing about the progress of a document dealing with library statistics. In addition, there were representatives of vendors there, and, it appeared, they were particularly interested in the electronic resources standard.

One discussion that cropped up was: Are these "standards" or "guidelines?" Standards used in engineering, networking, or technical areas are specifications; a product that does not meet a standard will probably not work with other equipment or software that do meet the standard. In other words, standards are necessary for interoperability of different products. When you hook up an Ethernet network card with Category 5e cable and hook up that cable to an Ethernet switch, it should all work together, no matter which vendor produced the various pieces because each adheres to the same standard. These specifications are included in purchase orders, so there is a legal component, too, and products not meeting the standard can be returned to the supplier.

Guidelines, on the other hand, are admonitions; and if your definition of a volume differs from another library's, the statistical equivalent of interoperability, comparability of the data, is violated. However, most likely, no one will ever know. We do not audit our reported data in the library world. I am with the guidelines crowd on this one because the **NISO Z39.7** standard is clearly more like a guideline than a standard.

The history of definitions used in collecting data about libraries is dealt with in **Robert V. Williams** "The Making of Statistics of National Scope..."² In it, he reviews the earlier history of a push for standards which began in the 19th century. His conclusion, which appeared in 1991:

Unfortunately, the appearance of an "official" national standard, even one with extensive cooperation across all the various library associations has not meant adherence to the definitions, procedures, and recommended list of variables in it...³

There have, of course, been areas of improvement, as he acknowledges, but these

standards have apparently not had much effect in the past on the course of library data. In effect, they are a top-down approach to improving library statistics and one done when there are not many worker bees to carry out the actual work.

I think, however, things are now different and this set of standards can lead to revolutionary changes, partially because of this standard, other work of **NISO**, and changing technology. There is cause for optimism, particularly with the measurement of electronic resources, although what lies ahead will take work.

Before outlining what I think is doable in Section 4, I want to discuss two things — first, important aspects of library data, which I discuss in Section 2, and data generally, which I discuss in Section 3.

2. Library data background⁴

Kendon Stubbs observed a dichotomy between two functions of data: "support" and "illumination."⁵ Support is the larger of the two and usually the more immediate. It comprises activities like decision support and justifying budgets. Illumination is the research and scholarship function, and the state of quantitative research in the field has been discussed by **Wallace**,⁶ **Williams**,⁷ and **Wyllis**.⁸ Briefly: we do not have much quantitative research, and what we do have is scattered over wide, disconnected areas.

This fact about quantitative research is an important one because it illustrates a fundamental problem with our data because we lack something many other disciplines have: a tradition of quantitative research. Data are a form of evidence and have strengths and weaknesses like any other form of evidence. Numbers will not answer every question, but they often can provide insight into those answers and eliminate some possibilities from consideration. Data are not magic, but a tool. The academic field of Statistics came into being when data analysis moved from the support to the illumination function and analysts started trying to understand what is going on behind the numbers, probed what caused what we see, and asked how to manipulate it.

The illumination function is not better in any sense, but only that it is different and necessary in a scientific field. In this sense, we librarians are pre-scientific in our use of data. Much good work is done, nonetheless. The **Library Research Service**,⁹ headed by **Keith Curry Lance**, provides useful services from data such as its "Fast Facts." The **Association of Research Libraries'** data series were developed largely for support but can also be used for illumination.¹⁰ These categories of support and illumination should be mutually reinforcing, and in healthy fields they are.

The library field is a large, cooperative field in which different people do different things. Not everyone will be inclined to work

with numbers any more than everyone would want to be a cataloger or work with acquiring electronic serials. The library field should, however, have some people who work with data in order to sustain a community of people who advance our understanding of our measures. I will return to this point shortly.

Adriano Balbi's 1835 *Statistical Essay on the Libraries of Vienna and the World* presages almost any discussion of library data we have today. Size of a library, he found, did not relate to a library's quality. There was a "stunning disparity of opinion" in various estimates of the various libraries he was trying to compare; in thinking about the reasons, he posited that:

[people] repeat sometimes without spirit of criticism the exaggerated evaluations coming from the mouth or the pen of an unconscientious librarian, who believes he will enhance the renown of the establishment which was entrusted to his care by exaggerating the number of its volumes...¹¹

He has more observations, but in our understanding of our data we are about where we were then. In 167 years, we have not learned much about our data.

So what? Well, data used for support are made better when there is an active illumination function. Scholars working with data naturally improve them. They investigate the error structure of the data, document data, and get

useful pictures of the quality of data in other ways, such as sampling and prototyping. Studies done with samples allow a picture of the state of the world without having to do a major study of all libraries, for instance; hence, they develop other techniques for measuring aspects of libraries. In the process, more is learned about the strengths and weaknesses of different variables and techniques for using these variables to answer questions.

What motivates the illumination function is getting a better understanding of the state of matters being investigated, and scholars have shown imagination in getting data for questions that interest them. And, too, when a disagreement on a point arises, the pressure to get and analyze data increases, including criticizing the data of one's opponents and attempting replication. The process results in examining data more carefully and arriving at an understanding of what the data in question are measuring and how. Such an enterprise makes data better through constant repetition, recursion, and replication.

After having seen data analysis elsewhere, this kind of improvement of data is, I think, almost instinctive and an inevitable result of systematic analysis. The **Association of Research Libraries (ARL)** collection of institutional data is clearly the premier library data series, reflecting, as it does, a systematic and improving series dating back to 1907/08 —

continued on page 87

The Charleston ADVISOR

Critical Reviews of Web Products for Information Professionals

Yes

ENTER MY SUBSCRIPTION FOR ONE YEAR

Name _____

Title _____

Organization _____

Address _____

City/State/Zip _____

Phone _____ Fax _____ Email _____

Signature _____

I am Interested in being a Reviewer. ☐ Yes

618 S. Monroe Way, Denver, CO 80209 Ph: 303-282-9706 Fax: 303-282-9743 <http://www.charlestonco.com>

Order Yours Today!

Libraries Pay Less than

\$6.⁰⁰ per review

With Publishers at Less than

\$10.⁰⁰ per review

HOW MUCH
ARE YOU
SPENDING?

Web Database and
Quarterly Print are now
available at the low
Introductory Price

\$295.⁰⁰ for Libraries

\$495.⁰⁰ for all others

Devil's Advocate from page 86

albeit one limited to a small collection of large libraries. The first issue of the predecessor **Gerould** series reported five variables ("Books in Library," "Added Last Year," "Spent for Books, Periodicals & Binding," "Number [of] Assistants," and "Total Salaries" for 14 academic libraries.¹² Over the years, **ARL** has added new libraries and also added variables as conditions changed or to answer questions that arose. It has dropped variables, too, as they have lost value or if they have not proved useful. In addition, the influence of this series has had echoes in other library series of data and the **Z39.7** standard.

As I mentioned, I have had the opportunity to work with data in two other areas: US banking and the data measuring information network traffic. Bank data have been collected in the US since the 1860s and with increasingly systematic methods. A large number of people work on these data and there are legal sanctions for misrepresenting the state of a bank on the regular quarterly filings of data on the condition of banks.¹³ There is active research taking place on these data and the data have an important support function and are used in illumination also. The literature on information network traffic is reviewed in **Molyneux and Williams**. This literature deals with a complicated data problem with huge sets of data but here, too, there is active research attempting to understand how networks work.¹⁴

It is quite striking to me how the library world differs from these other two in the use of data both to illuminate and to support. We seem to attract few technically inclined people to the field. That is, we seem to attract few who would have the inclination to analyze data and also too few, I feel, to manage the technical infrastructure that we are building in this changing information world. We do not have a critical mass of people with the inclination to do this work. The good news is we do not need many. In the librarian's house there are many mansions and this can be but one. I do not know how many it would take to sustain a self-sustaining research effort on our data. 10? 100? 1000? Whatever the number is, it is certainly more than we have. I have often thought such an effort could be jump started through concerted action.

Now a pep talk about data!

3. A few general observations on data

Oskar Morgenstern in his *On the Accuracy of Economic Observations*¹⁵ observed that people in one field often denigrate their own data while regarding data from other fields as "good." I often think of this observation when I find people denigrating pretty good library data, then using figures like the **Consumer Price Index (CPI)** as if it were an example of "good" data. If you think that about the **CPI**, I suggest you read **Morgenstern**. I also refer you to the essay by **William Kruskal**¹⁶ that deals with some of the vagaries of data — even good data. It is a nice and thoughtful article.

We understand the difficulty of defining a "volume" because we have thought about it and seen this definitional problem up close. When I show people outside the field the **Gerould/ARL** data, they can see its strengths: it runs for a long time — longer than the published historical series of bank data from the **Federal Deposit Insurance Corporation (FDIC)**,¹⁷ has relatively consistent definitions, is well documented, when analyzed shows reasonable results, and is available electronically. Someone else without the familiarity most of us have may think our data are not so bad.

I once posited that the result of the expulsion from the **Garden of Eden** was bad data.¹⁸ Before that unhappy event, I argued, data were good, easily obtainable, and reliable. Now, alas, it is a fact that data are difficult to obtain, that getting useful data takes hard work and seriousness of purpose, and that trustworthy and reliable data are not built in day. For example, the **FDIC's** effort is serious, involves a number of smart people, and if you lie to them, you could end up in jail.

I think such seriousness is also reflected in the census of **Caesar Augustus**. He wanted to know how many people were in the **Roman Empire** and he did not want local authorities, to paraphrase **Balbi**, "who believes he will enhance the renown of the establishment which was entrusted to his care by exaggerating the number of its" citizens and its wealth. Such a result could have had him committing the resources of the **Empire** incor-

continued on page 88

Endnotes

1. NISO Web page "About NISO." <http://www.niso.org/about/index.html>.
  2. Robert V. Williams, "The Making of Statistics of National Scope on American Libraries, 1836-1986: Purposes, Problems, and Issues." *Libraries and Culture*, v. 26, no. 2, pp. 464-485.
  3. Williams, p. 478.
  4. A number of the ideas in this section originally appeared in: Robert E. Molyneux, "More Hortatory than Factual: Fremont Rider's Exponential Growth Hypothesis and the Context of Exponentialism," in *For the Good of the Order: Essays in Honor of Edward G. Holley*, Delmus E. Williams et al. eds. (Greenwich, Connecticut: JAI Press, 1994), pp. 85-117, particularly 104-111.
  5. Kendon Stubbs, "On the ARL Library Index," *Minutes of the 108th Meeting, May 1-2, 1986* (Washington, DC: Association of Research Libraries, 1986), p. 19.
  6. Danny Wallace, "The Use of Statistical Methods in Library and Information Science," *Journal of the American Society for Information Science*, v. 34, pp. 402-410.
  7. Williams, "The Making of Statistics of National Scope...," p. 481.
  8. Ronald E. Wyllis, "Teaching Descriptive and Inferential Statistics in Library Schools," *Journal of Education for Librarianship*, v. 19, pp. 3-20.
  9. The Library Research Service's Web page is at: <http://www.lrs.org/>.
  10. The Association of Research Libraries has a number of statistical publications and an ongoing effort to improve its statistics. A listing of its statistical publications is at: <http://www.arl.org/pubscat/index.html#stats>. The oldest series is continued with annual publications of the *ARL Statistics*. An interactive version of the *ARL Statistics* is at: <http://fisher.lib.virginia.edu/arl/index.html>.
  11. Adriano Balbi, *A Statistical Essay of the Libraries of Vienna and the World*, translated by Larry Barr and Janet L. Barr. (Jefferson, North Carolina: McFarland, 1986). This is a translation of *Essai statistique sur les bibliotheques de Vienne* (Vienne, Frederic Volke: 1835). See pages 41-42 of the translation for this statement and his discussion of the vagaries of library data.
  12. Robert E. Molyneux, *The Gerould Statistics, 1907/08-1961/62* (Washington DC: Association of Research Libraries, 1986), frontispiece, page 4, and elsewhere. Or in the second edition on the Web at: <http://fisher.lib.virginia.edu/gerould/tgschap1.html>, section 1.3.2 "Variables Reported in the Gerould Data," and in passing elsewhere.
  13. Don Inscow, *Milestones in the Collection and Use of Data for Federal Deposit Insurance, 1934-2001*. (Washington, DC: Federal Deposit Insurance Corporation, 2001). Available on the Web at: <http://www2.fdic.gov/hsob/milestones.pdf>. I worked at the FDIC for Mr. Inscow in the Division of Research and Statistics and, among other things, helped put bank data on the Internet starting in 1993.
  14. Robert E. Molyneux and Robert V. Williams, "Measuring the Internet." *Annual Review of Information Science and Technology*, 34, 1999 (2001), pp. 287-339.
  15. Oskar Morgenstern, *On the Accuracy of Economic Observations*, (Princeton, New Jersey: Princeton University Press, 1963).
  16. William Kruskal, "Statistics in Society: Problems Unresolved and Unformulated," *Journal of the American Statistical Association*, v. 76, 1981, pp. 505-515.
  17. The FDIC's *Historical Statistics on Banking* starts in 1934. It is available at: <http://www2.fdic.gov/hsob/>.
  18. Molyneux, "More Hortatory than Factual...," p. 107.
  19. *Gospel According to Luke*, Chapter 2, verses 1-8.
  20. The World Wide Web Consortium's Web site is at: <http://www.w3.org/> and its XML Web pages begin at: <http://www.w3.org/XML/>.
  21. For instance, the US The House of Representatives has published the necessary documents to form an XML infrastructure for the legislative process. See: <http://xml.house.gov/>.
  22. Roy Tennant, *XML in Libraries* (Neal-Schuman Publishers: August, 2002). ISBN: 1555704433.
  23. Project Counter's Web site is at: <http://www.projectcounter.org/>.
  24. The ARL "Measures for Electronic Resources (E-Metrics)" Web page is at: <http://www.arl.org/stats/newmeas/emetrics/index.html>.
  25. International Coalition of Library Consortia (ICOLC). Web page is at: <http://www.library.yale.edu/consortia/>.
  26. Paul Metz and John Cosgriff, "Building a Comprehensive Serials Decision Database at Virginia Tech," *College and Research Libraries*, v. 61, 2000, pp. 324-334.
- Endnotes with URLs are also linked at: <http://www.molyneux.com/atg/>.

Devil's Advocate from page 87

rectly. What did he do to ensure the integrity of his count? He made every-

one return to "his own city" to be assured the count was controlled by his people.¹⁹ That is serious and his decision echoes through time. We sing "O Little Town of Bethlehem" and not "O Little Town of Nazareth" because Caesar wanted to make sure he got a good data. The continual work of the ARL over the years, similarly, shows a serious, long-term enterprise that continues to investigate new measures as an ongoing process and to update the technology of the data's use and dissemination.

The uncertainties surrounding the reported number of volumes held are pretty well known. After all, what is a "volume" and is a volume reported by university X the same as one reported by university Y? What about X in 1952 and 2002? In fact, volumes reportedly held are usually a sum of net volumes added plus a number handed down from one's predecessors. It is a number easy to criticize, however, when analyzed correctly, it yields reasonable results.

However, not all numbers are equally uncertain. Employees have names and tax records, and accountants can certify expenditure figures. We can usually be sure of these numbers, particularly after they are collected for a few years as there are often anomalous data in the earlier years of collecting a new variable. Now, while we can be sure of the numbers, what they mean is another question. We have historic numbers on the number of scrolls in the library at Alexandria, but who knows what to make of them?

Also, a part of the nature of data is that requirements for them change. We do not now collect variables we would like to have and the reality is that defining new variables, testing them, and collecting them well often takes effort and time. But the idea that there are good data and that they can be obtained by plucking them from the Tree of Knowledge is, sadly, not correct.

We are not alone in having problems with our data. That should not deter us from addressing those problems and we should be up to the seriousness of the task.

4. Z39.7

The Z39.7 standard deals with the "emetrics" variables at: <http://www.niso.org/emetrics/emetrics.cfm>.

Two observations:

1. This subset of the Z39.7 standards result from the considered opinion of a number of people and are the best guess of the NISO committee that worked on these data about which data items are of most use in measuring electronic resources. While these measures would all be useful to have, they are sparse in the matter of use of those resources. NISO and others have discussed eJournals, but many that would be of use to the matter of use of electronic resources are not included in this standard.

2. Other NISO standards have the infrastructure for eXtensible Markup Language (XML), as Patricia Harris, Executive Director of NISO, pointed out at the meeting, while Z39.7 does not. Actually, the ARL data and related series have electronic formats for transferring data and they work pretty well but these formats are not readily adaptable to electronic serials use.

XML is a general markup language initially developed by the World Wide Web Consortium²⁰ to overcome difficulties with the HyperText Markup Language (HTML), which does not share XML's extensibility. XML can be adapted to a large variety of functions; central to our story here is that XML allows, among other things, the creation of data exchange formats for specialized purposes and is being used in industry and government²¹ for that purpose. Increasingly, XML is being used in libraries, as demonstrated by many Websites and Roy Tennant's newly published *XML in Libraries*.²²

In order to use XML, one must create a formal definition of the format, either a Document Type Definition (DTD) or an XML Schema. Then parties in the exchange have to agree to use it. Finally, programs to manipulate these data will need to be written.

5. With these facts in mind, I want to focus on a support question many of you are familiar with.

The question of the use of electronic resources is of importance to readers of *ATG*. Shrinking budgets and large costs of electronic resources

continued on page 89

is causing an immediate problem of gathering data to support decisions about how best to use these expensive electronic resources. Librarians are flying blind.

There are a number of people who have written and discussed aspects of the question of measurement of electronic library materials. **Project Counter (Counting Online Usage of Networked Electronic Resources)**,²³ **ARL's eMetrics project**,²⁴ and the **International Coalition of Library Consortia (ICOLC)**²⁵ are all working on aspects of deciding which variables to collect and how to report them. These efforts involved librarians and vendors. From the presence of vendors in these efforts and at the meeting at **ALA**, it occurred to me the vendors might be interested in prototyping a data exchange format to be used after the variables are decided on.

Everyone needs more data, and the vendors are trying to find a means of making the data available through these efforts. But, a word of caution. Two problems will not be changed by an agreement between the vendors and librarians. First, these are the early days of the process of finding which variables are of most use, and so these data have not gone through the iterative process I discussed earlier that is necessary to improve data. People want data now and they are impatient to get data about electronic resources but good data are not built in a day. Second, is the matter of format. How will we merge all these data from the different vendors in the future — even if we all agree on what data we need and they can supply?

I believe the formatting question presents an opportunity to start the creation of a data exchange format using **XML** to facilitate the interchange of data dealing with electronic resources, and it would be complementary to **NISO's** work and the work of these other groups, perhaps a bottom-up development that can prototype an **XML** library data exchange format for this specific area.

Arriving at the format and putting the data in it would require effort, as would developing the infrastructure to make the format useful to librarians.

Currently, as the reader is aware, librarians get reports on electronic resources from different vendors in a variety of non-compatible formats that cannot be easily merged. If an **XML** format were in widespread use, the data would be in a compatible and easily merged format.

Then what? If appropriate utilities were written to take the reports from the vendors and "parse" them and "transform" them, in the language of **XML**,

they could be merged and put directly into whatever format was desired by the librarian, say, a spreadsheet or a Web page or **PowerPoint**. It would make the support function more useful to have different data from different vendors in one report.

If an interested group of people had similar support data, it would spur an illuminating conversation in how best to use them and the data would get the scrutiny necessary to improve data, as has happened with the ongoing development of the **ARL** data.

Such a prototype, though, would only be a step in the journey. These resulting data would be new data dealing with a new subject and subject to the problems all such

new data have. The data are not a magic bullet to solve everyone's problems, but with work and time, the data will improve. For instance, **Paul Metz** and **John Cosgriff**²⁶ describe a method for working with the difficulties with the data they had. The article presents a transparent methodology for their comparisons of the various non-comparable data items they had. If you disagree with their weightings, you can adjust them to your purposes.

I will be working on a prototype and will be contacting some of you for data. I hope to present a preliminary report at the **Charleston Conference 2002**.

Desperately Seeking Copyright — Digital Object Identifiers: Just an Idea or an Innovation of Value?

by **Edward Collieran** (Director, Publisher Relations, Copyright Clearance Center, Inc.) <ecollieran@copyright.com>

The global groundswell for the adoption of yet another numbering system would appear to be underwhelming at best. After all, we already have **ISBNs**, **URLs**, **EAN/UPCs** and a plethora of other numbers. Why, indeed, should we invent another?

But an unusual convergence of interests among librarians, researchers and publishers has, in fact, already propelled the new **Digital Object Identifier (DOI)** in a few short years from an abstract concept to an operational numbering system for identifying and locating intellectual property in digital format.

DOIs have rapidly claimed a major role in the **STM (Scientific, Technical and Medical)** publishing world and are poised to venture forth to other publishing sectors and media formats as well. The first four

DOI registration agencies already have issued five million **DOI** numbers identifying mil-

lions of individual articles in 6,500 journals. In addition, the **Library of Congress National Digital Library Program** and the **Defense Advanced Research Projects Agency (DARPA)** virtual library for the **Defense Department** use the same underlying technology. Recent announcements include a major Asian adoption and a large-scale multinational European **DOI**

initiative. And a dynamic plug-in for **Adobe's Acrobat** has been developed which brings the power of **DOI** to pdf files. That's a lot of traction in only two and a half years!

What exactly is a **DOI**? **DOIs** are unique, permanent numbers assigned to specific objects (for example: digital articles, images, music or video), which remain unchanged (in effect, move with them from Website to Website, or publisher to publisher). For a simple analogy, **DOIs** are like cell phones which owners take with them wherever they go. **URLs**, in contrast, are like home or office phones in fixed locations; they are of limited usefulness when you are not at that location. Note that **DOIs** themselves are not an encryption or locking technology; they merely identify digital content as registered and provide a mechanism for the publisher to offer multiple viewing and payment options.

DOIs are enormously helpful to researchers, especially in hyperlinked footnotes, because they transport researchers directly to the article, irrespective of its location, rather than driving them to Websites that no longer exist. **DOIs** also are helpful to those librarians who need to track and control the materials that researchers purchase.

The **DOI** has two components. The prefix is the naming authority, assigned to a publisher, for example; and the suffix — added by the publisher — identifies the specific article, image or media as well as its metadata (that is, data about data, such as the underly-

continued on page 90