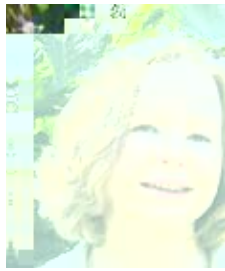# From the Chair

Data 101

Amy West

As many of you probably know, I spent my first seven years at the University of Minnesota as an official, card-carrying government publications librarian. Then, in 2007, the libraries did a bit of reorganizing and decided it was time to formalize the work I'd been doing and now I'm a card-carrying data librarian. Of course, because so much of the data floating around in the world is generated by governments or through government funding, I remain deeply interested in public access to government information. Thus, I've been thrilled by the surge of interest in "government data" during the last year. I've also been a bit frustrated by just how broadly data has been construed without any routine acknowledgment of the many flavors of data that are of such interest. Conveniently, while I'm GODORT Chair, I have this column in which to discuss various elements of government data and what data might mean to government information specialists with respect to dissemination, access, formatting, usability, and preservation of increasingly large quantities of government information at all levels. This first installment will cover some of the primary meanings of "data."

First, there's data that we're familiar with—numeric data files representing censuses, surveys, geospatial coordinates, and sensor measurements. These files may be true data, in that they represent content at the level of observation, or they may be summary tables or statistical visualizations generated from the numeric data. Typically, data is excluded from depository programs, but summary tables and statistical visualizations regularly included, either on their own or as part of larger textual publications.

Key to this conceptualization of data is that it is separate from text. Text equals publications equals the rows and rows of books and microfiche in libraries. However, from the perspective of a computer programmer, data is anything that may be structured. Anything. For example, in H.R. 1105, "bulk data download" referred to availability of legislative branch textual material in bulk and in a structured form.[1] This is a perfectly acceptable use of data, but it's not the one typically used by government information librarians. More importantly, the government has finite resources. If efforts to get agencies to expend their energies on the development of structured textual data are successful, then it will most likely be at the expense of the traditional publication. This isn't a bad thing, but it is most definitely a different thing. Data as a way of structuring text, rather than something other than text, will most likely have profound effects on what it is that libraries collect from governments, whether through a depository program or not.

Academic libraries are talking about data primarily in terms of the biomedical, physical, and natural sciences. Data from these disciplines are often produced on very large scales and have massive storage, description, analytical, and preservation challenges. However, equally important is the relationship of federal funding to research, the government's definition of research data, and the role of copyright. The reason that so many academic libraries have become so interested in scientific data is that, in addition to being an opportunity to create new kinds of collections while participating in the advancement of information management, that's where the action is. According to the National Science Foundation, the federal government expended more than twenty-two billion dollars in fiscal year 2005 on research in science and engineering at universities and colleges in the United States.[2] Thus, many academic libraries are now looking at the research data resulting from federally funded research as a potential target for collections. However, in this context, "research data" is not only different from what any librarian might expect, but it's especially different from what a government information librarian would expect because, unlike typical government publications, copyright is an issue.

In OMB Circular A-110, research data is defined as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings . . ."[3] Effectively, research data is defined as a typical journal article, with (possibly) some supporting data. Further, "the recipient may copyright any work that is subject to copyright and was

---

GODORT Needs a Fresh Logo!

Submit your design by December 1, 2009. Guidelines are available at wikis.ala.org/godort/index.php/GODORT_Logo_Contest.

From the Chair

developed, or for which ownership was purchased, under an award."