

Editorial

Bringing Librarianship to E-Science



In venues where I have been recently, the topic of e-science often comes up, and comments range widely. Some colleagues ask—why should we librarians become involved in data management at all? Others argue—we have a professional responsibility to help meet the challenge that scientists face in managing massive amounts of data.

Beginning in the 1950s, scientists and engineers were among the first to use computers to enhance the research process. They moved from research in the laboratory to modeling that uses algorithms, computer programs, and new information technology. The use of modeling helped to inform, enhance, and speed up the research process. However, modeling requires and generates massive amounts of data.

Today, data sets are scattered around the world in individual researchers' computers, university centers and labs, and national and/or disciplinary digital repositories. Data sets have been generated that contain great detail about water and other environmental conditions, the atmosphere and the heavens, the mysteries of plants and the human body, and the most elemental, physical particles, currently the focus of nanotechnology.

Numerical data sets have become the lifeblood of computational research in many scientific and engineering areas. Today's researcher, working independently or in a team, determines what data are needed to undertake a specific computational test. The researcher then shares the findings from the test with colleagues through prescribed, vetted methods—a research journal or a conference paper. We as librarians have this

last step—management of the formal literature that has evolved over the last 150 years—well in hand, but not the earlier phases in the life cycle of scientific information that have become so important today. We should learn from our past success.

During the first half of the nineteenth century, researchers shared their findings by sending their papers to colleagues in order to share findings and receive comments. To facilitate sharing, researchers came together and eventually founded professional societies including, among others, the American Society of Civil Engineers (1853), the American Chemical Society (1876), the American Society of Mechanical Engineers (1880), the Geological Society of America (1888), the American Physical Society (1899), and the American Astronomical Society (1899). It wasn't long before researchers needed an additional method for communicating research results beyond the professional meeting. Consequently the professional journal was born. The challenge then became to acquire, organize, and preserve these research publications for continuing reference and referral. The proliferation of these research publications created a need for a "repository" to house, organize, and provide access.

Toward the end of the nineteenth century, academic and research libraries accepted responsibility to collect and house research publications. Prior to this time, academic libraries were generally small, housing static collections that primarily supported the teaching of the classics. So as scientific research grew, academic libraries grew as well.

To help researchers locate material in the chaotically growing body of material

published by scientists and others, three developments occurred at the end of the nineteenth and the beginning of the twentieth century. What we call “knowledge management” today started as classification/subject headings, cataloging, and indexing.

First, through the collaboration of librarians with disciplinary subject specialists, the Library of Congress developed a classification scheme and a subject-heading system that brought materials related in content together, both on the shelf and in the catalog. The result was a consistent, uniform, and universal access-and-retrieval system. (Melville Dewey also developed the Decimal Classification System, which could be substituted for the classification part of the duo.)

Second, in order to provide a researcher or student with a consistent method of discovery and access, the *Catalogue Code* was agreed upon in 1908 by the American Library Association and the Library Association. Uniform bibliographic description ensured that books and journals in library collections would be described consistently no matter which library held them.

The third development came from outside libraries. The publishing industry responded to the need of researchers for ways to locate articles within research journals. In the early twentieth century, indexes and abstracts were started, including *Science Abstracts* (1903) and *Chemical Abstracts* (1907).

So why have I taken us down memory lane? In the twenty-first century, the challenge of data-set management is similar to the one faced nearly 150 years ago by researchers who wanted to share their findings among colleagues. Discoverability and availability must come to data management to foster today’s scientific research. A logical, accepted, intuitive structure must be developed to facilitate discovery and access to data sets throughout the world. Libraries must step up and assume responsibility for archiving data that underlies the research article. Those who argue that it is not the role of libraries to

archive data used in the scientific research process should remember this: massive quantities of original documents held by archives in libraries throughout the world are collected, organized, and preserved, for the most part as, these documents “raw data” until a researcher in the humanities or social sciences uses them to answer a research question. (Questionable: are human readable documents in archives really comparable to bits of data? Want to modify or revise this statement? How does that really differ from the collection and use of scientific digital data?)

The National Science Foundation (NSF) and the Association of Research Libraries (ARL) have recognized the role librarians can play in data management. For an in-depth discussion of the challenges and opportunities, read the following publications:

US National Science Board. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*. September 2005. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_1.pdf

Association of Research Libraries. *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering. A report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: the Role of Academic Libraries in the Digital Data Universe*. September 26-27, 2006, Arlington, VA. <http://www.arl.org/bm~doc/digdatarpt.pdf>

The application of library science principles and methodologies, such as cataloging, classification, and resource sharing, can be reinterpreted to meet the specific needs of scientific digital data management and described in terms that are more expansive and expressive of today’s challenges, such as *metadata*, *taxonomy*, and *open source*. In this way a network of well-documented data sets can be built that will facilitate the retrieval of data by researchers, today and into the future. Who else is better qualified than we librarians to bring this about?

James L. Mullins
Purdue University