THE CHANGING ROLE OF LIBRARIES IN SUPPORT OF RESEARCH DATA ACTIVITIES

A PUBLIC SYMPOSIUM

BOARD ON RESEARCH DATA AND INFORMATION

On the afternoon of June 3, 2010, the Board on Research Data and Information (BRDI) organized a public symposium on *The Changing Role of Libraries in Support of Research Data Activities* in Washington, DC. This event featured presentations by senior managers from three federal library organizations and from the Association of Research Libraries, who have examined the role of libraries for research data in some depth and whose members are institutions with a great deal at stake in this area. The symposium concluded with comments by two Board members, a university professor and researcher working in the data-intensive field of geographic information and a university professor of information science. The symposium was moderated by Prof. Michael Lesk, chair of the Board.

The Role of Libraries in Digital Data Preservation and Access – The Library of Congress Experience

Deanna Marcum, Library of Congress (www.loc.gov)

The Library of Congress (LoC) has been collecting publications in all formats and in all subject areas (except in medicine and agriculture) since the 1800s. It has vast historical collections in science and technology, including papers of noted scientists such as Newton, Copernicus, Darwin, Pavlov, and Einstein. The collections include works of scientific and technological historiography and secondary works of analysis and interpretation. Everything that comes through the copyright system is offered to the Library of Congress. In the 20th century, the LoC began to focus on topics important to Congress including material that supports research on current issues, legislation, public policy, and emerging technologies in areas such as alternative energy, cyber-security, and climate change.

In looking ahead, the LoC has decided to develop a strategy for digital information and data related to sciences and technology that is appropriate for a national library. The LoC created an *e*Science Team, headed by Peter Young, which includes fifteen staff from across the institution. The group meets regularly to determine the Library's role in digital science. The *e*Science Team approach to discussing this role fully and adequately requires outside perspectives and talking with experts at federal science agencies about scientific data, digital data curation, and their use. The Team is developing pilot projects so that we can see what is required. For example, the LoC has been collecting a lot of geospatial data and transferring them to a pilot project so staff can investigate and analyze them as possible models for the Library's role. The Team formulated study questions regarding target digital data sets to determine the challenges. To develop a deeper understanding of the Library's role, the pilot project will clarify workflows, policies, and technical issues on transfer, ingest, management, and access related to digital data sets.

The Library of Congress staff was eager to participate in BRDI because the Board is focused on the improvement of management, policy, and use of digital data and information for science and the broader society, and the Library's interest is in supporting researcher investigations from its national collections. In rethinking its collections, the LoC needs to be sure that congressional members and staff have the most up-to-date and complete information as they create new policies. The old way of dividing collections by subject matter or discipline is not necessarily going to work today. We need to move from the traditional way of organization to a new way to organize and make accessible digital science materials for data-driven research and interdisciplinary research.

The LoC faces many challenges. The Library needs ways to engage communities of practice in developing preservation, archival and access policies, methods, and tools. It needs a new digital infrastructure to address long-term uses of digital data sets. It needs to be able to support this new way of working and new procedures have to be in place. The Library's initial efforts in this regard include the National Digital Information Infrastructure and Preservation Program (NDIIPP). The NDIPP's purpose is to develop a national strategy to collect, preserve, and make available significant digital content, especially information that is created in digital form only, for current and future generations. Essentially, the Library has to move from the reading room to the research center, which is to say that it needs to move from traditional ways to a more open environment that allows experimentation and collaboration with experts at the LoC. This will require sustained access and the continual use and reuse of data. The new requirements are collaboration, cross-institutional approaches, infrastructure, and the integration of collections. The Library of Congress therefore is working on ways to bring it all together so people can find what they need, whatever it may be.

More Data, More Use, Less Lead Time: Scientific Data Activities at the National Library of Medicine

Betsy Humphreys, National Library of Medicine (www.nlm.nih.gov)

The National Library of Medicine (NLM) has large collections in data categories that include substances, sequences, clinical research, and related taxonomies, nomenclature, and ontologies. The NLM's biggest challenge is how to handle more data coming in more rapidly. There is a great investment in research that generates data. We need to find out what the data are, how to organize them, and better ways to standardize them. Obviously, as the price of generating and storing data goes down, we get more data. The exponential increase in data input is huge. Examples of how much data input is growing include PubChem, ClinicalTrials.gov, and the UMLS Metathesaurus.

The primary strengths of the National Library of Medicine are a believable commitment to the effective organization of data and a permanent, robust infrastructure with strong partnerships, international collaborations, and heavy use. Systems that get used get better. The NLM has connections between different kinds of data and information. Its weaknesses are limited resources and, as a result, less user outreach and training than is desirable.

The National Library of Medicine began creating and storing digital bibliographic information in the 1960s, and its first scientific data database was built in the 1970s. The National Center for Biotechnology (NCBI) was created in 1988. The NCBI exists to design, develop, implement, and manage automated systems for collection, storage, retrieval, analysis, and dissemination of knowledge concerning molecular biology, biochemistry, and genetics. It performs research into advanced methods of computer-based information processing capable of representing and analyzing the vast number of biologically important molecules and compounds. It enables persons engaged in biotechnology research and medical care to use these systems and methods. It also coordinates, as much as is practicable, efforts to gather biotechnology information on an international basis and to connect all kinds of data.

In terms of international collaborations, the International Nucleotide Sequence Database Collaboration is a premier example, with regional centers in the United States (GenBank), Europe (EMBL) and Japan (DNA Databank). This is a great model in the sense that it provides backup and access to data. This is international data sharing at its best. More data and more connections mean it will get used more, and this escalation of use means more data input.

The NLM's current activities and future plans include continued emphasis on improving the quality of the input, such as tagging, standardization, and explicit links made at the source of the data. The focus is on increasing data curation efficiency and promoting standards and best practices, U.S. partnerships and international collaborations, computer center security and efficiency, and better discovery, retrieval, and display methods. Because of the escalating input of data, we need to improve the quality of the original input so it does not require so much curation on our end. The NLM also needs to increase curation efficiency and to create better retrieval and display methods for people who use the data.

Libraries in the New Research Environment

Joyce Ray, Institute of Museum and Library Services (www.imls.gov)

The Institute of Museum and Library Services (IMLS) was created in 1996, and thus its entire history has been in the digital age. The IMLS is the primary source of federal support for the nation's 123,000 libraries and 17,500 museums. Its mission is to create strong libraries and museums that connect people to information and ideas, and to help build the capacity of libraries and museums through grant-making, convenings, research, and publications. With that mandate, the online knowledge universe presents a number of challenges. These include building the online content landscape, addressing copyright barriers, enhancing discovery and creating tools to support advanced research and re-use of content, and developing sustainable repositories at the right economies of scale.

The IMLS funds grants that address these issues through research, development, and demonstration projects, primarily in its National Leadership Grants Program. This program promotes innovation, impact, and collaboration in one or more of the following areas: advancing digital resources, demonstration, research, and library-museum collaboration. The IMLS awarded its first National Leadership grants in 1998 and began funding digitization projects in that first year. From the beginning, the intention of digitization grants has been to fund projects that help to identify best practices for digitization and managing digital data. IMLS grants have led to the development of statewide digital collaborations, many involving museums and archives as well as libraries, in over 40 states, plus thematic collections, as well as best practice guides such as the Framework of Guidance for Building Good Digital Collections (http://framework.niso.org/). IMLS grants have also addressed issues such as aggregation of digital content, metadata, interoperability, copyright management, the development of discovery and presentation tools, sustainable repositories, and digital curation. Since 2002, IMLS has funded a research and development project at the University of Illinois, Urbana Champaign (UIUC) that is aggregating content using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), the same protocol that is being used to build Europeana, the European Union's digital library. The UIUC project has now become the largest online curated resource for American history, with over 900 collections and more than one million item-level records. The project has also provided a testbed for research on a range of technical issues relating to aggregation, normalization, and discovery of data (see http://imlsdcc.grainger.uiuc.edu/history/).

Data from big science are relatively easy to handle. Data from small science are more diverse and difficult to manage. How should such data be preserved and managed? What are the best practices for digital curation, particularly for small and medium sized repositories that could have relevance to any number of repository data managers? The IMLS has supported research, led by the Purdue University Libraries Distributed Data Curation Center, to develop data curation profiles to answer the questions of who is willing to share data, when, and with whom for a range of scientific disciplines and sub-disciplines (see http://d2c2.lib.purdue.edu/). The profiles will help the Purdue Libraries to begin providing data management services to researchers in those disciplines that are most open to sharing their data. A subsequent grant is enabling the project team, in cooperation with partners at UIUC, to distribute a toolkit and conduct

workshops for librarians in other institutions so that they can create additional curation profiles and establish their own data management services.

The final challenge is who is going to do this work? The IMLS 21st Century Librarian program provides funds to prepare the next generation of library and information science (LIS) educators and workforce. Since 2006, a major focus of this initiative has been support for the development of digital curation programs in graduate schools of library and information science. Masters' and doctoral programs in digital curation have been developed with IMLS funding at UIUC, the University of North Carolina at Chapel Hill, the University of Tennessee, and the University of Syracuse, and masters' level programs have been established at additional schools.

Supporting *e*-Science: Progress at Research Institutions and Their Libraries

Karla Strieb, Association of Research Libraries (www.arl.org)

The Association of Research Libraries (ARL) includes membership of 125 research libraries in the United States and Canada. Affiliate organizations include the Coalition for Networked Information (CNI) and the Scholarly Publishing and Academic Resources Coalition (SPARC).

The ARL created an E-Science Task Force in 2006 and hosted an NSF-funded Workshop: *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering* (see http://www.arl.org/bm~doc/digdatarpt.pdf). The 2007 Task Force Report recommendations are at http://www.arl.org/bm~doc/ARL_EScience_final.pdf, and the proceedings from the 2008 ARL/CNI Forum, *Reinventing Science Librarianship* are here: http://www.arl.org/resources/pubs/fallforumproceedings/forum08proceedings.shtml. The Task Force took stock on what should be the role of the research library and came to realize that libraries need to operate in a multi-institutional, international, collaborative context. What is the library's role from creation of data all the way through to the use? What is needed are well-integrated models.

In the summer of 2009, the ARL surveyed its membership to understand the status of their institutional planning and the emerging roles of research libraries in advancing *e*-science. The survey looked at what libraries are doing and how they fit into what research institutions are doing to advance *e*-science. Already, three-fourths had infrastructure in place or planned. Almost all reported both central and decentralized activities, usually in a hybrid patchwork. The survey revealed a wide range of emerging practices, roles, and partnerships and surfaced a broad set of model resources created by respondents.

ARL member libraries reported various pressure points. Organizationally, pressure points included limited recognition of the importance of *e*-science support and coordination issues concerning providing support for *e*-science within and across institutions. Resource pressure points included finding staff with relevant expertise, technology infrastructure, and budget constraints. There are few business models for aligning the demand for library services with research funding. Budgetary constraints are not just about not having enough money; they raise questions about how to create new kinds of models and how research libraries can develop models that tie service demand to funding.

The ARL survey produced many other interesting facts about the *e*-science activities of ARL members. To learn more about the survey, visit http://www.arl.org/rtl/eresearch/escien/esciensurvey/.

Comments by Michael Goodchild, Professor of Geography

University of California, Santa Barbara

Spatially referenced data are the primary data in the geosciences, as well as being important in many other fields, such as the biosciences and the social, behavioral, and economic sciences. Framework or foundational data answer the question, "Where am I?" One of the great achievements over the past few years is interoperability, making it easy to convert from street names to longitude and latitude, and so on. Web 2.0, social networks, and other online innovations have already influenced the geospatial world tremendously. A spectacular example is OpenStreetMap, a project begun by a graduate student at University College London, which has expanded to a world-wide effort. This year, soon after the Haiti earthquake, the community involved in OpenStreetMap produced a very high quality map of Port-au-Prince, which has become the official map of the UN's response effort.

Where does this leave the university library? The UC, Santa Barbara has a large map library that has two main functions: (1) to have basic coverage of the globe, and (2) to cover the local area. There is no need for all map libraries to have identical coverage of the globe, but rather they should concentrate on a more regional and local specialization. My university library is slowly transitioning to a more regionally-based collection. Many institutions of higher learning now have some version of a geographic information system (GIS) or spatial data center, recognizing that dealing with this type of data is not simply a matter of accession, custodianship and circulation, but also requires a large amount of assistance and expertise.

What does that mean for long-term archiving? A spatial data center probably has less longevity than the library, so we have been trying to emphasize that the preservation and archiving of the data should still be with the library. More general issues include time, a critical aspect, and conservation principles. In the geospatial world, time is critical. Current, real-time data is collected, and continuous monitoring has become more important. We are more concerned with images of the world now as opposed to sometime in the past. The earthquake in Haiti is an example of this critical aspect. A sample map was produced within days and possibly hours of the earthquake, showing enormous destruction and activity. Conservation principles mean that we also need to consider not just what is changing fast, but what is staying the same. Unless we understand this, we cannot understand the future.

There has been an exponential increase of material flooding into the libraries, but there is so much no one can keep up with it all. An example of this is that my ability to write has increased, but my ability to read has stayed the same and perhaps has even diminished. Yet the times to attain a Ph.D., to educate, or to publish have not changed much.

One characteristic of *e*-science is that it is data rich and heavily based in simulation. It is also potentially fast. Traditionally, a library contained materials that were static. The budget, staff, and physical space all stayed largely the same. What would a real-time library look like? Would it be a library that is focused on *e*-science, rapid science, and continuous monitoring? Events are not just characterized by volume, but also by speed, so we would have to have catalogs that update themselves very rapidly.

Why data matter to librarians – and how to educate the next generation

Christine Borgman, University of California, Los Angeles

I have been reflecting on prior speakers and there were several repeated themes: the librarian must be a partner with the scientist, starting planning early in the data life cycle is crucial, early engagement is key to taking the temporal approach, and reconfiguring institutions from being a reading room to a full research center is increasingly important. In looking at the distribution and content of data-related courses being taught in the United States, we found that the course scope did not start early enough in the scholarly process. We drew upon curricula elsewhere, but developed most of our own new material for the first course offering at UCLA. Entitled Data, Data Practices, and Data Curation, the instructional objectives are these:
Students will learn

1. To distinguish between the many forms of data, how data vary by scholarly discipline, and how they are used throughout the scholarly life cycle;
2. The roles that data play in research collaborations;
3. To distinguish among different types of data collections, repositories, and services;
4. Basic principles of public policies for data;
5. A basic knowledge of data curation practices in the library and archive fields; and
6. Professional criteria for selecting and appraising data.

The course began with an overview of data, data practices, and data curation, e.g., “the big picture,” and tried to alternate between guest speakers and ourselves. One thing we learned was that if you really want to find out what is going on with the data, you have to talk with doctoral students, not just the professors. When the students leave a project, it can be very hard to reconstruct the data.

The term concluded with the realization that we were just getting the conversation started. Next year (2010-11) we have extended the scope to a two-course sequence (winter-spring). The winter term course will focus on the first four of these learning objectives and the second course (for which the first is prerequisite) will focus on the latter two objectives.

Symposium Agenda

The Role of Libraries in Digital Data Preservation and Access - The Library of Congress Experience
Deanna Marcum, Library of Congress

More Data, More Use, Less Lead Time: Scientific Data Activities at the National Library of Medicine
Betsy Humphreys, National Library of Medicine

Libraries in the New Research Environment
Joyce Ray, Institute for Museum and Library Services

Supporting E-science: Progress at Research Institutions and Their Libraries
Karla Strieb, Association of Research Libraries

Comments by Michael Goodchild, Professor of Geography
Michael Goodchild, University of California, Santa Barbara

Why data matter to librarians – and how to educate the next generation
Christine Borgman, University of California, Los Angeles

Panel discussion of the invited speakers and Board members, and general discussion with the audience.