Ann Campion Riley

Data management and curation
Professional development for librarians needed

Professional development is one of the core purposes of ACRL. New areas of knowledge within librarianship are obvious choices of subjects for professional development programming. Within scholarly communication, the most recent service area is data management and curation. At its Midwinter meeting, the ACRL Board of Directors affirmed its commitment to providing professional development in data management and curation through its decision to fund new instruction and training opportunities in this growing area.

Most information technology professionals think in terms of months or perhaps two or three years as the timeframe for archiving data. Librarians think in terms of 200 years or longer. Research libraries in particular have demonstrated this mind-set by keeping printed materials indefinitely, and participating in programs to enable them to share in long-term preservation and access. Solutions like the Center for Research Libraries that offer offsite storage for less-used books and journal volumes are well known.

The challenges of preserving data for the long term are well known now, too, with the obvious problem of recovering data from outdated tools like floppy disks in the early days of personal computing. Much work has been done in the last ten years on the preservation and migration of information and images from the last century, much of it by groups like the Internet Archive, National Digital Stewardship Alliance (the Library of Congress's preservation directive), and many other libraries and agencies.

Many researchers have developed limited and local ways to store their own data and datasets they create or use. Development of discipline-based repositories such as Dryad,[1] GenBank,[2] and others have been an important step in data preservation. ICPSR[3] at the University of Michigan is a pioneer in data curation for social sciences research. Its funding model, with a high level of staffing, offers a standard most of the newer efforts in data preservation cannot reach. The New England Collaborative Data Management Curriculum,[4] with a focus on health science data, is another pioneering initiative. Many of the current leaders in the ACRL efforts have worked with that curriculum and have used it to guide their educational programs.

Another star in this area is the Purdue University Libraries' program in data curation. The creation of the Purdue University Research Repository (PURR)[5] models the centrality of the role libraries can play in offering these services to their parent institutions. Staff at PURR are available to help researchers with all aspects of organizing and storing research data. Librarians as well as technical staff offer support services. Subject specialists who can work with

Ann Campion Riley is acting director of libraries at the University of Missouri and ACRL President, email: rileyac@missouri.edu

the researchers are another important group. Data storage options vary widely based on the type and size of datasets, and so do metadata consultation and creation. Consultation with researchers at the beginning of their projects, before data creation begins, is the optimal time to develop data structure and metadata format. The prominence of PURR and the promotion of its services help that to occur.

After the ACRL preconference Getting Down to Brass Tacks: Practical Approaches for Developing Data Management Services offered in Portland, Oregon, in spring 2015, the Digital Curation Interest Group,[6] led by past-convener Yasmeen Shorish, proposed several options for ACRL to continue to develop offerings related to the topic. The ACRL Board chose to fund the creation of a curriculum for a program patterned on the Scholarly Communication Roadshows.[7] These roadshows have been offered for about seven years and have been very popular. Roadshows are one-day events hosted regionally by local institutions. The local institutions provide venues and logistical assistance, and the presenters, selected and compensated by ACRL, provide the program. In the past, some local presenters were featured on related topics in the afternoon.

Some affinity groups, such as the Research Data Alliance and the Medical Library Association, have of course already been offering training in data management and curation, and the ACRL leadership certainly does not wish to duplicate them. However, the need is widespread and growing, and many librarians are unable to attend specialized conferences. The opportunities for interested people need to be wide and varied. Because of the complex nature of the topic, along with differing levels of need, the market can absorb many offerings.

The growth in this area has been significant, as demonstrated by a proliferation of conference presentations and articles on relevant topics. At a recent presentation I gave on the role of libraries in providing data management and curation services, I read from the Research Data Access and Preservation (RDAP)[8] program the titles of the presenters. These included data librarian, digital curator, digital services

specialist, library data manager, digital librarian, digital data librarian, and many other similar titles. Clearly, the profession does not have a consensus on what to call colleagues who work in this area. Similarly, some of the qualifications for these positions include library training, but many do not, having rather an emphasis on one of the subject areas served.

At my own institution, the University of Missouri, we have discussed the need for people to translate between the data creators and the technical workers, plus those charged with the preservation of the data. In our planning we have used the term *data concierge* to describe the sort of activities these staff will need to undertake. In the hospitality industry, a concierge typically fields a wide variety of questions from hotel guests in an unfamiliar location who need information on where to shop, eat, and find services.

Researchers who produce data often find themselves in an unfamiliar territory when they need to organize and store the data and are faced with choices when they may not know all possible options. Analyzing needs and then analyzing what the researchers are creating in various formats is a crucial part of providing data management and curation services, as well as directing researchers to possible storage solutions. These data concierges need not be any one type of staff. They may be metadata specialists, subject specialists, or research generalists, but they need a specific body of knowledge, with training in local and nonlocal options for data curation, to help researchers effectively.

Newer modes of teaching have made data use and data literacy important issues at colleges without a major research mission. No one is surprised to hear that large research-intensive institutions, where faculty and graduate students work on large, often federally funded grants from the National Science Foundation (NSF) and the National Institutes of Health (NIH), have a need for data curation. Many may be surprised to learn that in an introduction to psychology class, again at my own institution, in addition to participating in small experiments, some students are now required to enter

and, at times, learn to interpret data from those same experiments as part of their educational experience. As computing pervades learning on many campuses, the capacity to do research in automated and analytical ways has changed how researchers work.

Federal mandates for management and sharing of research data are not brand-new developments. As of January 2016, it will be five years since NSF began requiring data management plans with grant applications, and almost 13 years since NIH began implementing its data sharing policy.

Anecdotally, researchers report that grant applications are now being screened more intensively on the details of data management plans than in the past, with specific questions on the feasibility and lifespan of proposed solutions. As a part of librarians' role in preserving scholarship, offering researchers help to find places and ways to preserve the data in and from their research is an essential extension of the ways librarians have for many years helped authors find places to publish, and then have kept the books and journals that have resulted.

While certain federal agencies require providing public access to the results of their funded research, we should not conflate data management and curation with the Open Access Movement. Much research data produced cannot and should not be made publicly available, because the data is proprietary to a privately funded corporation or the data has national security issues attached to it. Unrestricted access to data should remain the default, though, and librarians need to encourage researchers to follow that path, even when the data is not the result of federally funded research.

Related to data access is the key issue of data governance. In my own work I have found that when asked the question of who owns the data they produce, almost half of researchers give wrong answers. Data governance policies vary by funder and by institution. Those offering services to researchers on storing and preserving data need at least to make researchers aware of the issues. Copyright issues may also be murky, regarding specifically curated

datasets related to articles published in non-open access journals.

ACRL's Research and Scholarly Environment Committee (ReSEC)[9] is one of the goal area committees of the organization, and has done much work to educate librarians on open access and data management issues. The Scholarly Communication Toolkit,[10] an online resource managed by ReSEC, gives practical guidance and provides examples and case studies for reference.[11] Because transforming scholarly communication is one of the three primary goals in ACRL's current strategic plan, the Plan for Excellence, supporting the work of ReSEC has high priority. The successful program of Scholarly Communication Roadshows offers a model of how ACRL may deliver data management and curation training in the future to our members, training that will benefit many people working in libraries in this growing area of service to our user communities.

Notes

1. Dryad, http://datadryad.org/.

2. GenBank, https://www.ncbi.nlm.nih.gov/genbank/.

3. ICPSR, www.icpsr.umich.edu/.

4. New England Collaborative Data Management Curriculum, http://library.umassmed.edu/necdmc/index.

5. Purdue University Research Repository, https://purr.purdue.edu/.

6. Digital Curation Interest Group, www.ala.org/acrl/aboutacrl/directoryofleadership/interestgroups/acr-igdc.

7. Scholarly Communication Roadshow, www.ala.org/acrl/issues/scholcomm/roadshow.

8. Research Data Access and Preservation Summit, www.asis.org/rdap/.

9. Research and Scholarly Environment Committee, www.ala.org/acrl/aboutacrl/directoryofleadership/committees/acr-schlcomm.

10. ACRL Scholarly Communication Toolkit, http://acrl.ala.org/scholcomm/.

11. Examples and case studies of scholarly communication initiatives, http://acrl.ala.org/scholcomm/?page_id=448. ⌀