Research Data Services: A New Focus for Librarians

Carol Tenopir, University of Tennessee, Knoxville, Tennessee, USA

ctenopir@utk.edu

CONCERT Proceedings, 2013 Electronic Resources and Consortia

Common wisdom says that science has entered a "fourth paradigm" that is more collaborative, more computational, and more data intensive (Hey, Tansley, & Tolle, 2009a) than the previous experimental, theoretical, and computational paradigms. This emerging scientific paradigm is often referred to as e-science or e-research (Hey, Tansley, & Tolle, 2009b). Increased reliance on technology in all parts of scientific endeavor, or cyberinfrastructure, and the establishment of data management and data sharing mandates by many research funding bodies have motivated academic libraries to respond to the changing needs of their faculty and student researchers and consider how best to engage in e-science through the development of library-based research data services (RDS).

The need for sharing and reusing data and the need for better data services comes from several motivations. The first of these is the growing open science movement, with government initiatives to make data publicly available. In the U.S., data management and data sharing mandates have been established by the National Science Foundation, the National Endowment for the Humanities, the National Institutes of Health and others (University Libraries, University of Minnesota, 2011). There are mandates for sharing the data that is collected as the result of government grants or other government funding in many countries today, including the U.S., United Kingdom, and Australia.

Merely instituting a mandate for sharing does not ensure that data will be properly stored and preserved so it is usable now and in the future, however. Both a motivator for data management and preservation and a reason for acting now is the vulnerability of much of the data and datasets that are collected by researchers. A majority of scientists store their data on their own hard drives or in their offices. Backups may be on USB drives or on paper printouts. Thus, data is vulnerable to many disasters, including:

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies
- Format obsolescence
- Legal encumbrance
- Human error

- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations and requirements (Michener et al., 2012)

The United States National Science Foundation established the DataNet program to create exemplar partners to address "…one of the major challenges of this scientific generation: how to develop the new methods, management structures and technologies to manage the diversity, size, and complexity of current and future data sets and data streams". (NSF, 2007)

One of the first DataNet projects, of which I am a part, was DataONE (Observational Network for Earth) that concentrates on preserving and making accessible data and datasets in the earth and environmental sciences. DataONE's vision is to "Provide universal access to data about life on earth and the environment that sustains it, as well as the tools needed by researchers." (ORNL, 2013).

DataONE is not merely a system, it is a combination of people, hardware, and software. It strives to build a community and culture of data management and data sharing within the earth and environmental science community and provide a robust system for depositing, accessing, preserving, and using data.

DataONE has four key principles:

1. Data should be part of the permanent scholarly record and requires long-term stewardship.

2. Sharing and reuse maximize the value of data to environmental science.

3. Science is best served by an open and inclusive global community.

4. Evidence-based assessment is necessary for practice and governance in the dynamic data environment.

The DataNet program is based on recognizing the need to integrate "library and archival sciences, cyberinfrastructure, computer & information sciences, and domain science expertise to provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline" (NSF, 2007).

Libraries have a natural leadership role for data services, because they are one place in their institution that sees the broad picture across all constituents or subject disciplines. Libraries and librarians are accustomed to working with multiple collaborators across an institution. They facilitate interdisciplinary work, information access, and knowledge acquisition through a variety of collections and services and can take a leadership role in a variety of research data services.

Research data services are services across the full data lifecycle, including data management planning, digital curation (selection, preservation, maintenance, and archiving), and metadata creation and conversion. Services can be technical/hands-on or informational/consultative (Tenopir, Sandusky, Allard, & Birch, 2013).

To understand the full range of possible research data services, it is important to understand the research data lifecycle. There are many versions of data lifecycles; the one used by DataONE is shown here.

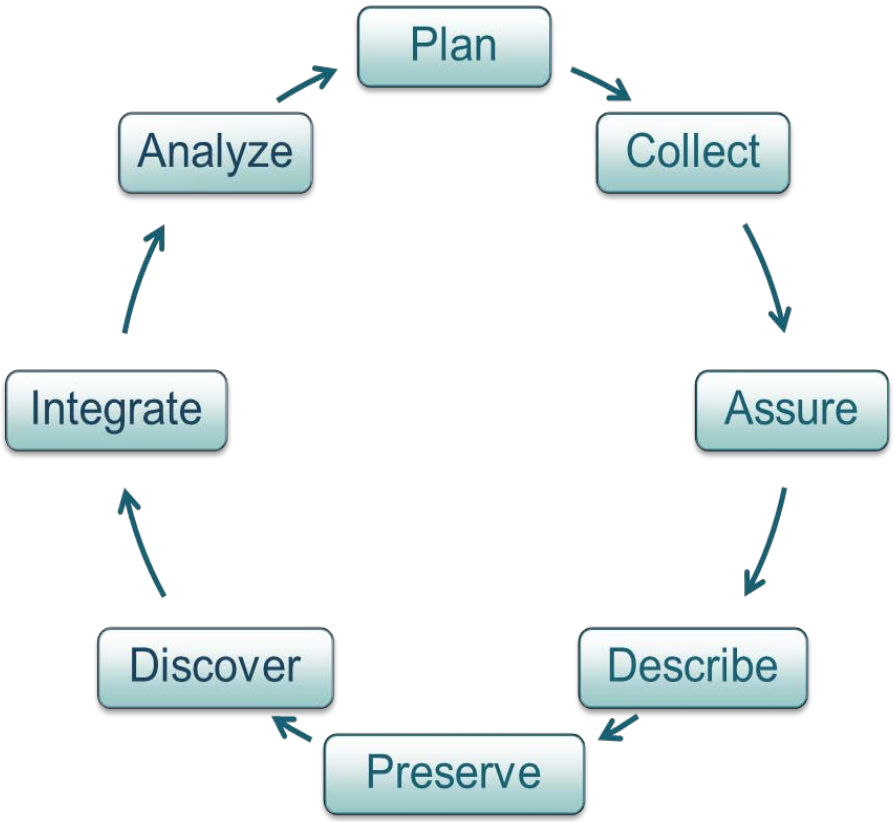

Figure 1. Research Data Lifecycle (DataONE)

The daily work of scientists is focused on only some of the steps in the lifecycle—notably on planning, collecting, and analyzing their data (and somewhat on assuring through quality control.) Describing data through metadata, preserving data in secure and long-term fashion, organizing data for discovery or knowing where to discover others' data, and reusing and integrating data collected by others are often not yet a part of scientists' normal routine. All of these areas are natural places in the lifecycle for librarians to assist through research data services.

At this early stage, many librarians may have questions about their role in research data services at all stages of the data lifecycle (Figure 2.)

Figure 2. Librarian and Library Challenges in the Research Data Lifecycle

Librarians may wonder:

- What is the priority of RDS in my institution?
- How involved should I be (or can I be) with helping researchers formulate metadata?
- Is there a repository that will accept my researchers' data or do we need to have an institutional repository for data?
- What will be the library's role in selecting and deselecting data if we have a repository?
- What role should we have in preserving data for both the short-term and long-term?
- What can librarians do to help researchers discover data?
- Should I be directly involved in research teams as a data expert?
- Do I have the education, experience, and ability to provide RDS?

Technical or hands-on services that a library may consider include (Tenopir, Birch, & Allard, 2012):

- Providing tech support for RDS systems/data repositories
- Deselecting datasets to be removed from a repository
- Preparing datasets for deposit
- Creating or converting metadata
- Identifying datasets for deposit in repositories

Consultative/informational data services that a library may consider include (Tenopir et al., 2012):

- Consulting on data management plans (DMPs)
- Data and metadata standards consulting
- Collaborating with other RDS providers on campus
- Assisting with finding and citing data
- Developing web guides and finding aids
- Directly participating on a project
- Discussing RDS with others on campus
- Training co-workers on data issues or providing training opportunities for librarians

The DataONE Usability and Assessment working group conducts initial and subsequent evaluation of data practices, needs, and attitudes of various stakeholder communities that could benefit from the DataONE system. These include scientists, librarians, decision and policy makers, publishers, educators, data managers, and citizen scientists. Baseline assessment surveys of scientists, academic libraries, and academic librarians help reveal current practices and opportunities for the future for the library role with research data services.

The baseline assessment of scientists, published in PlosONE in 2011, provides insights into how libraries can help with the data management barriers faced by researchers (Tenopir et al., 2011). Over 1300 researchers from academic, government, and other institutions responded to a survey distributed worldwide, mostly via email, with an embedded link to the questionnaire in the body of the email. (For more details see Tenopir et al., 2011.)

Several results suggest natural areas for librarians to help, for example, when asked about what metadata standard scientists use to describe their data, answers ranged across a wide variety of schema, with no one schema being used by more than a small percentage of respondents (see Table 1). By far the most common answer was “none”, that is the respondents use no metadata (and perhaps do not even know about the concept.) Librarians have multiple roles here,

educating researchers about existing metadata standards in their field, helping to design or make available metadata templates or tools, or taking on the task of adding and editing metadata for their researchers' data. (Tenopir et al., 2011)

Metadata Standard	**Percent Use**
DC (Dublin Core)	2.2%
DwC (Darwin Core)	1.7%
DIF (Directory Interchange Format)	1.0%
EML (Ecological Metadata Language)	7.9%
FGDC (Federal Geographic Data Committee)	7.9%
ISO (International Standards Organization)	8.0%
OGIS (Open GIS)	8.0%
Other	6.8%
Metadata standardized within my lab	22.1%
None	56.1%

Table 1. Metadata Standards Used by Scientists (Tenopir et al., 2011)

There is also a gap between "willingness to share data" and the reality of making data easily accessible to others (Michener et al., 2012). In the scientists' survey, although 75% of scientists agree, "I share my data", only 36% agree, "Others can access my data easily". Also, the willingness to share may have its limits for a variety of reasons. More than three-quarters of respondents are willing to place "some" or their data into a central repository, but less than half (41%) are will to place "all" of their data there (Tenopir et al., 2011). Helping to differentiate between datasets that are ready and appropriate for sharing and those that are not or that need to go into a dark archive for preservation purposes but not for sharing, is another way that librarians can assist their scientist constituents.

There are many reasons why data is not currently being made available, as shown in Figure 3. Although libraries cannot help with all of these, taking the top three barriers from researchers of insufficient time, lack of funding, and no known place to put data could be a natural effort of libraries.

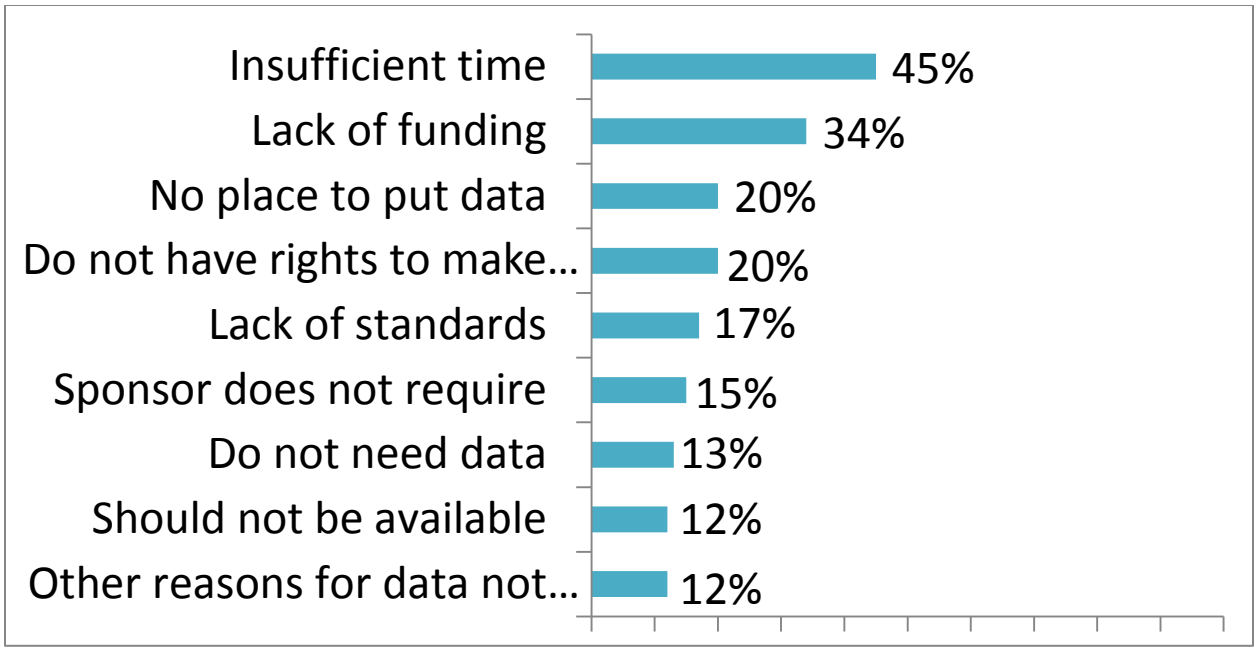

Figure 3. Reasons why scientists do not make their data electronically available (Tenopir et al., 2011).

With these practices and needs of scientists in mind, surveys of academic libraries and individual librarians who work in academic research libraries, provide insights into opportunities for libraries in the context of the current state of library and librarian involvement with RDS and plans for the future.

Librarians who work in academic research university libraries were asked about their background, preparation, and opinions regarding RDS. Over three hundred North American librarians, who are most likely to be involved with RDS (science librarians, data librarians, technology librarians, etc.) responded. At the same time, a survey of 223 North American academic library directors measured official library policy, services offered by the library, and plans for the future. Comparing opinions of individual librarians with the libraries they work in provides interesting contrasts. (Tenopir et al., 2013).

Individual librarians responding to the survey had a variety of experience with RDS. In response to the question: “Do you interact with faculty, students, or staff in support of their research data services (RDS) as part of your regular job responsibilities?” 27.9% said RDS was integral to their job, 40.5% said they have occasional RDS responsibilities, and 31.5% have no RDS responsibilities. Attitudes towards the importance of RDS differ significantly among these three groups. In response to the statement “RDS are as important as other services”, 82% of the integral group, 68% of the occasional group, and only 32% of the no group agreed. In response to the statement “RDS are a priority at my library”, 67% of the integral group, 40% of the occasional group, and only 19% of the no group agreed. (Tenopir et al., 2013)

The integral group is also much more likely to agree that "I have the skills, knowledge, and training necessary to provide RDS", "I have sufficient subject expertise to provide RDS to my patrons", "My library provides opportunities to develop skills related to RDS", and "My library supports me to attend conferences/workshops on RDS." If RDS is to expand in libraries, opportunities for education, workshops, and skills development need to be offered to a broader range of librarians. (Tenopir et al., 2013)

The number one motivation for those librarians who are not currently providing RDS to do so, is "if my patrons request" such services, followed by "if RDS becomes a responsibility in my job", and "if my institution becomes more involved with RDS". With a growing focus on research data sharing from government mandates and international expectations, the first and third of these statements is likely to occur. (Tenopir et al., 2013)

According to Tenopir et al. (2012): "Currently, a minority of US and Canadian academic libraries are offering research data services, with more planning to begin in the next year to two years. More libraries are offering or planning to offer informational/consultative-type services, rather than technical assistance services." The most commonly offered or planned service is reference services to locate data or datasets, followed by creating webguides or other finding aids for locating data. In the future more libraries plan to provide research data services, specifically reference-type services, than are currently providing them. (Tenopir et al., 2012)

Not surprisingly, RDS services of either the reference type or technical type are more likely to be offered now or in the future by libraries in PhD-granting universities rather than baccalaureate or two-year/associates academic institutions. (Figures 4 and 5.)

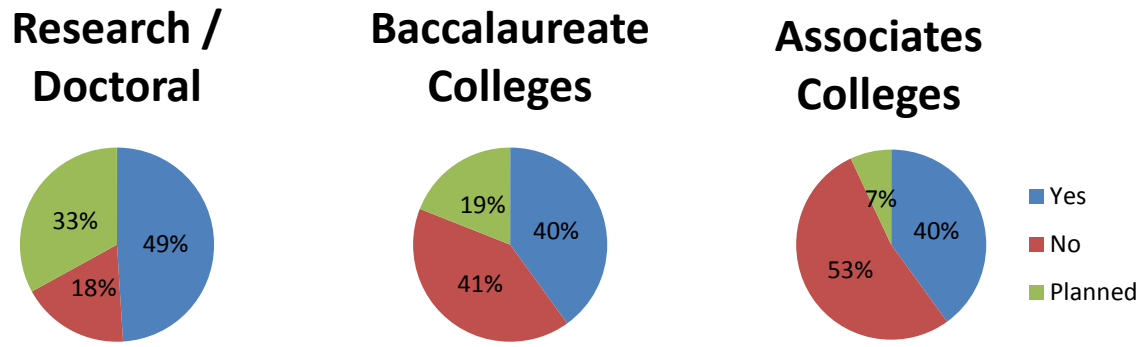

Figure 4.Reference support services offered by type of institution (Tenopir et al., 2012).

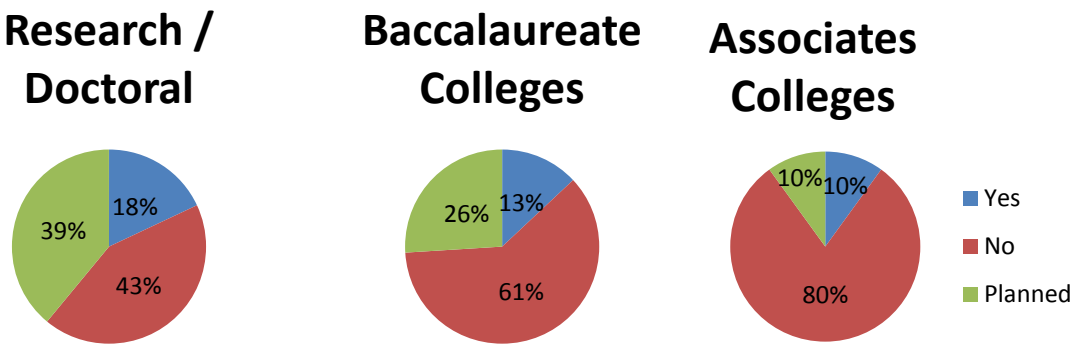

Figure 5.Technical data services offered by type of institution (Tenopir et al., 2012).

In contrast to the replies of the librarians, only about one-third (31%) of PhD institutional libraries offer support for staff education and training, and only 17% of baccalaureate institution libraries and 16% of associates/two-year colleges. (Tenopir et al., 2012)

Clearly, additional educational opportunities for librarians to learn about a variety of RDS are needed. Data science programs are being developed both within the information sciences curriculum and science curriculum. These efforts will help educate a new generation of data services librarians. In the meantime, continuing education and workshops for existing librarians who hope to upgrade their skills are important.

RDS are important and in line with libraries' missions and roles; however, libraries are at an early point in the movement to an RDS focus. Libraries are resetting priorities, realigning responsibilities, and creating opportunities to develop skills. A new generation of data services librarians are being educated (and are finding jobs), but other librarians need opportunities to develop RDS skills. (Tenopir et al., 2013)

According to a report for the Association of Research Libraries: "This is next-generation librarianship. The curation of research data is an activity that has gained traction in the wake of library and information science programs offering concentrations in data curation and institutes in digital curation, promising a cohort of librarians qualified to meet the challenges of managing data." (Hswe & Holt, 2010)

References

DataONE. (n.d.). *Best Practices.* Retrieved from http://www.dataone.org/best-practices

Hey, T., Tansley S., & Tolle K. (2009a). *The fourth paradigm: Data-intensive scientific discovery.* Redmond, WA: Microsoft Corporation. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf. Accessed 2012 Nov 2

Hey, T., Tansley, S., & Tolle, K. (2009b). Jim Gray on eScience: A transformed scientific method. In T. Hey, S. Tansley, & K. Tolle. (Eds.), *The fourth paradigm: Data-intensive scientific discovery.* Redmond, WA: Microsoft Corporation. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf. Accessed 2012 Nov 2

Hswe, P. & Holt, A. (2010). *NSF Data Sharing Policy*. Retrieved from http://www.arl.org/focus-areas/e-research/data-access-management-and-sharing/nsf-data-sharing-policy

Michener, W.K., Allard, S., Budden, A., Cook, R.B., Douglass, K., Frame, M., Vieglais, D.A. (2012, September) Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. *Ecological Informatics*, *11*, 5-15. http://dx.doi.org/10.1016/j.ecoinf.2011.08.007

National Science Foundation. (2007). *Sustainable Digital Data Preservation and Access Network Partners (DataNet) – Program solicitation NSF 07-601*. Retrieved September 30, 2013 from http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141&org=CISE

Oak Ridge National Laboratory. (2013, April 30). *Supercomputing and Computation – DataONE*. Retrieved from http://www.ornl.gov/science-discovery/supercomputing-and-computation/projects/dataone

Tenopir, C., Allard, S., Douglass, K,, Aydinoglu, A., Wu, L., Read, E., Manoff, M., & Frame, M. (2011) Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE 6*(6): e21101. doi:10.1371/journal.pone.0021101

Tenopir, C., Birch, B., & Allard, S. (2012). *Academic libraries and research data services: Current practices and plans for the future.* Association of College and Research Libraries White Paper. Retrieved January 13, 2013 from http://www.ala.org/acrl/issues/whitepapers

Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. (2013). Academic librarians and research data services: Preparation and attitudes. *IFLA Journal, 39*(1), 70-78 Retrieved from http://www.ifla.org/publications/ifla-journal

University Libraries, University of Minnesota. (2011). *Funding agency and data management guidelines.* Retrieved November 15, 2012 from https://www.lib.umn.edu/datamanagement/funding