At the Watershed: Preparing for Research Data Management and Stewardship at the University of Minnesota Libraries

Leslie M. Delserone

Abstract

Like many research universities, the University of Minnesota (UMN), and its Libraries, attempts to understand the nature and intensity of data produced by its researchers, and address the management and stewardship of its institutional research output. The recent activities of the Libraries, in support of and in conjunction with campus-wide efforts, illustrate a set of approaches that show the Libraries to be at the watershed, integrated into and contributing leadership to the growing river of campus-wide exploration and planning of cyberinfrastructure needs. This brief article highlights the hiring of the "science librarian cohort," the Libraries' study concerning the needs and behaviors of scientific researchers, the implementation of the University Digital Conservancy, the Libraries' involvement in the UMN's Research Cyberinfrastructure Alliance, and its initiation of the e-Science and Data Services Collaborative.

Introduction

In Minnesota, the state of 10,000 lakes, watersheds matter, in part because a quarter of the state's watersheds drain into the Mississippi River. The systematic, institutionally driven[1] management of research data at the University of Minnesota (UMN)—and the UMN Libraries' involvement in this process—is approaching a watershed moment. The Libraries' ongoing work springs from the confluence of several sources and contributes to a river of activity at UMN concerning cyberinfrastructure needs.

At both national and international levels, the alarms sound about the "data deluge" in the sciences (Hey & Trefethen, 2003; Microsoft Research, 2004). In the United States, the National Science Foundation (NSF) explores a national cyberinfrastructure network, and includes considerations

LIBRARY TRENDS, Vol. 57, No. 2, Fall 2008 ("Institutional Repositories: Current State and Future," edited by Sarah L. Shreeves and Melissa H. Cragin), pp. 202–210

of data management and stewardship (National Cyberinfrastructure Council, 2006). The NSF requires the description of a data management plan in grant proposals (National Science Board, 2005),[2] and both NSF and the U.S. National Institutes of Health require researchers to share their data.[3] As libraries enter the conversation,[4] there is growing recognition, within the library profession and the scientific community as well, that our combined knowledge and skills may be valuable to this enterprise. Concurrent with these developments, the UMN set a challenging institutional goal: to be one of the top three public research universities in the world.[5] The Libraries' contribution to UMN's strategic positioning includes explorations of its role in the management, preservation, and stewardship of research data,[6] which some may see as an appropriate extension of the Libraries' archival role. The UMN Libraries' activities over the last few years exemplify one approach to exploring an academic research library's role in the management of research data generated at its home institution.

At the Headwaters

Within a watershed, several small and independent streams may contribute to river's formation. For the UMN Libraries, three activities—hiring, research, and program development—which may on the surface appear to be independent, converge to support a potential role in data management and stewardship at the UMN.

Science Librarian Cohort and Data Services Librarian

In late 2006, the Libraries hired three science librarians (the "science cohort"),[7] with one FTE each based in the biomedical sciences, the agbiosciences, and the physical sciences and engineering.[8] While these librarians retained some familiar duties—collection development; reference; liaison work with assigned departments and research centers—they also began to investigate interdisciplinary research collaborations occurring within and between various research arenas. In their respective domains, these librarians gained an understanding of the nature of data produced by their researchers, and of any data management requirements, either discipline-driven or mandated by funding agencies. As stated in the *Compact for the University Libraries, 2006–2007,* the science cohort would

> document emerging needs, particularly in e-science, for expert support and information and technology infrastructure, and working with OIT [the UMN Office of Information Technology] and the Libraries' Digital Development Lab, develop prototype online environments of content and tools to facilitate easier and more productive exploitation of the research literature and data resources.[9] (University of Minnesota Libraries, 2006, p. 5)

The UMN Libraries' "E-Science and Data Services Collaborative (EDSC)," described below, now encompasses and extends the work of

the science cohort librarians. Additionally, the data services librarian, appointed in 2007, has a leading responsibility in shaping services, which define the library's role in helping to ingest, manage, preserve, and make data resources—specifically social sciences data—accessible and useable. The librarian in this position will also develop programs to enhance access and information literacy with respect to data and statistical resources. Working collaboratively, the data services librarian provides a focus for expertise and coordination of services and collections.[10] For the UMN Libraries, a commitment to hires in newly-described positions makes available additional personnel in the sciences and social sciences, who will contribute to research and action (e.g., recruitment of data collections, development of grant proposals) in data services.

The "Science Assessment" Study

Shortly before the science cohort librarians arrived at Minnesota, the Libraries embarked on a systematic, internally funded research project, "Understanding Research Behaviors, Information Resources, and Service Needs of Scientists and Graduate Students: A Study by the University of Minnesota Libraries,"[11] otherwise known as the "Science Assessment" study. From October 2006 through June 2007, sixteen science librarians, including the science cohort, met with more than seventy faculty, postdoctoral, and graduate student researchers from the Academic Health Center, the Colleges of Biological Sciences and of Food, Agricultural and Natural Resource Sciences, and the Institute of Technology (UMN's College of Engineering and Physical Sciences). The scientists self-selected for participation by responding to invitations extended by the project coordinator or the liaison librarians with whom these researchers have close working relationships. Librarians participated as advisors, as focus group moderators or interviewers, and/or with data analysis and report development. The topics covered during the discussions ranged widely, from issues of discovery and access to the challenges presented by interdisciplinary collaboration.

For the purposes of this article, the most relevant information volunteered by the researchers came in response to the question: "If you seek assistance from the library, what kinds of help are you looking for? What kind of assistance is needed? (For grants? Publishing? Data curation and preservation?)" (University of Minnesota Libraries, 2007, p. 30). The scientists' comments could be grouped into three main themes: (1) data organization and manipulation, (2) data storage/security/sharing, and (3) data stewardship.[12] Several of those interviewed were looking for the Libraries' help in data organization and manipulation:

> The Libraries have a tremendous opportunity to lead this [teaching data organization skills to scientists], and to provide ways to interpret, validate, and build on the data produced. (Center for Library Initiatives, 2008, slide 6)

> There are multiple ways to retrieve things now, but why not data? . . . You could find pieces that are related to it, but still keep things organized by project, table, and descriptor. It would be great if you could create new folders with a common link, keep it in multiple formats or reassemble it. (Center for Library Initiatives, 2008, slide 7)
>
> If I'm taking raw data—sequencing work of genome—can I Google the data to find out what's known about it? Is there technology that will help me do it? How do you do raw data comparisons? Are there search engines just for data sets, even the ones that are constantly changing? (Center for Library Initiatives, 2008, slide 8)

On the subject of data storage/security/sharing, there were expressions of frustration over the lack of standards or procedures to reference; in some cases, scientists looked to the funding agencies for this information, while others expected the UMN to provide guidelines or define mandatory behaviors.

The data stewardship comments covered a wide range of opinion. Some felt that it is unnecessary:

> Am I worried it [data] won't be there in 20 years? No. Am I worried it won't be there in 100? It doesn't matter. By that point, data become irrelevant except as historical curiosity. (Center for Library Initiatives, 2008, slide 13)

One researcher confessed to not having really considered the need for preservation of his data; he had not thought that these data would be of future interest:

> It's important to maintain data for two or three years—saved on disks—but after that the field moves so quickly that it's no longer relevant…I hadn't really thought much about [researchers who might be interested in the work in 10, 20 or 30 years]. But it wouldn't be good if they couldn't find the data, would it? (Center for Library Initiatives, 2008, slide 14)

Several researchers commented that an individual scientist can make the decision:

> Data storage [preservation] is fundamental to all of us, but it's not as though there is an IRS rule for keeping it for 7 years. We keep data long enough for people to know about it. (Center for Library Initiatives, 2008, slide 15)

Completing the connection between data organization needs and potential data stewardship needs, a professor stated:

> The Libraries could facilitate the curation and preservation of data by scholars, and teach researchers how to better organize it (Center for Library Initiatives, 2008, slide 18).

For the UMN Libraries, focused discussions between knowledgeable science liaisons and the researchers they serve provides strong anecdotal evidence for the continuing exploration of the provision of data services.

University Digital Conservancy

An additional foundation piece toward data stewardship came online officially in late August 2007, with the launch of the University Digital Conservancy (UDC),[13] the University of Minnesota's institutional digital repository, sponsored by the UMN Libraries and the Office of Information Technology. The UDC, while an excellent repository for many materials including digital documents and images, is currently not the best platform for data stewardship. A visitor can discover a few static datasets (flat files, no supporting applications for manipulation)[14] in the UDC at this time, but librarians have not systematically recruited data. Discussions are underway concerning an alternative infrastructure for data deposits (e.g., Fedora), with the UDC home page serving as the "front door" to a variety of collections. The EDSC plans to investigate the data stewardship possibilities of the UDC in its current DSpace instance, as well as standards for data stewardship (e.g., metadata). Given the UMN Libraries' role in stewardship of the UMN research and activity record, there is interest in continuing an assessment of the UDC as a data repository. Having the digital institutional repository in place is a critical component of any institutionally-driven progress in data stewardship.

On the River

Given the confluence of the three examples mentioned above, additional activities at the UMN contribute to a river of activity on data management and stewardship. This section will present two efforts, one UMN-wide and the other Libraries-driven, that are relevant to data concerns.

Research Cyberinfrastructure Alliance

The considerations involved in data management, storage, sharing, and stewardship are not limited to discussion within the UMN Libraries. In fall 2007, the Libraries, the Office of Information Technology, and the Office of the Vice President for Research joined together to form the Research Cyberinfrastructure Alliance (RCA), and engage in a university-wide discussion and planning process, along with the Academic Health Center, the Institute of Technology, and the Colleges of Liberal Arts, Biological Sciences, and Food, Agricultural, and Natural Resource Sciences.[15] The RCA's work ties back to the University's strategic positioning, its goal being "to position the University to enable computationally intensive, interdisciplinary research for the 21st century" (University of Minnesota Research Cyberinfrastructure Alliance, 2008b, slide 5), and also to "align [the University's research efforts] with NSF" (University of Minnesota Research Cyberinfrastructure Alliance, 2008b, slide 6). As part of the RCA's environmental scan of the types of computationally intensive research in progress at the UMN, representatives from four laboratories discussed data management and other infrastructure concerns. [16]

> The interviews reveal researchers eager to work with campus partners to relieve themselves of the day-to-day burden of administering data management solutions. . . . Core needs of data storage and expert assistance are similar enough that a common solution may be feasible, especially if it is layered in a way that encourages the development of some domain expertise and long term relationships between researchers and those supporting their work. (University of Minnesota Research Cyberinfrastructure Alliance, 2008a)

Exploration of existing university services continues, and with the identification of gaps, the RCA will develop a plan for coordinated services. Where do the UMN Libraries fit in the RCA discussions? As a key member of the group, the Libraries turn attention to a new application of its ongoing work—selection, acquisition, and preservation, but now inclusive of data—as well as identifying and developing environments for interdisciplinary research, and leveraging liaison librarians' relationships with their faculty researchers. The Libraries' EDSC likely will contribute content as well as Web design expertise to the RCA's website, its public face to the campus community. Ideally, the Collaborative will

> research and define a clear UMN Libraries' role within the RCA [via]
>
> - data stewardship/archiving consulting and technology,
> - educational roles,
> - build[ing] scientist portals (similar to the Cornell VIVO model)
> - data management policy,
> - data repository certification (e.g., TRAC certification for data repositories), (EDSC Project List, unpublished working document).

E-Science and Data Services Collaborative (EDSC)

Simultaneous with its involvement in the campus-wide RCA, the UMN Libraries continues its own exploration of varied aspects of data management and stewardship. The Libraries have a collaborative working-group structure that is successful for research and taking action. Each collaboration consists of appointed groups of librarians and other staff, and addresses such matters as scholarly communication (University of Minnesota, Scholarly Communication, 2008), reference services (University of Minnesota, Academic Programs, 2008, September 29), information literacy (University of Minnesota, Academic Programs, 2008, October 13), and diversity outreach (University of Minnesota, Academic Programs, 2008, October 15). The EDSC (University of Minnesota, E-Science and Data Services Collaborative, 2008), initiated in spring 2008 with a ten librarian team, lists as its objectives:

- to build knowledge and capacity within the Libraries to support e-science and data services, leveraging existing expertise, projects and tools;
- to define our core services and areas of expertise in "data services" in the context of other campus services and initiatives such as the Minnesota Supercomputing Institute [which reports to the Office of the Vice President for Research, an RCA member];

- to define a potential new model for library liaison roles across campus that supports interdisciplinary science (including relevant social sciences);
- to contribute to University discussions about interdisciplinary research and teaching and to develop a framework for educating the campus about data policies, including those that support open data initiatives.

As of fall 2008, the EDSC team identified several potential projects, and will begin their efforts with educational outreach to library colleagues and the wider UMN community. In speaking with colleagues, the EDSC membership will consider a variety of topics, including:

> data curation policies of various major funding agencies, creation of data management plans by individual researchers, communication of existing data support services at the UMN [from an inventory conducted by the RCA], standard data submission procedures for selected data centers, similarities and differences in data archiving issues and resources for "big science," "small science" and social sciences/humanities research projects (EDSC Project List, unpublished working document).

In educating UMN researchers, the EDSC recognizes that the Libraries cannot be the sole source of information and outreach. It is necessary to partner with others in the RCA (e.g., the Office of the Vice President for Research) to develop and sustain the delivery of a data management and stewardship class (EDSC Project List, unpublished working document). Other potential projects emphasize data management policy—again, a collaborative venture; collaborating with colleagues involved in similar explorations and actions at other research institutions; and assisting UMN liaisons in discussions with their researchers about data management and stewardship, along the lines of the assistance provided by the "Scholarly Communication Collaborative" toward author's rights education.

River's End?

Looking back, one can identify events at the headwaters where the UMN Libraries' involvement began: hiring additional science librarians to support interdisciplinary collaborations between researchers, and among librarian colleagues; an environmental scan of the needs of UMN science researchers; and the launch of an institutional digital repository. But at the time of this submission, there is no Gulf of Mexico, as there is for the Mississippi River, to mark the endpoint of the Libraries'—or the University's—river of efforts and decisions. As this article goes to press, the work of the RCA and the EDSC is in its early stages. But perhaps the river of decision making and action cut by the UMN and its Libraries illustrate a potential path that other institutions may find useful to study or emulate.

Notes

1. In contrast to data management that is driven by a particular discipline's requirements (e.g., deposit of nucleic acid sequence in GenBank), the concerns of an individual researcher, and/or the requirements of a specific funding agency.
2. An example of an explicit request by NSF for a data management plan is this International Polar Year solicitation letter, available at http://www.nsf.gov/pubs/2007/nsf07008/nsf07008.jsp.
3. "NSF's policy position on data is straightforward: all science and engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected and preserved" (NSF Cyberinfrastructure Council, 2006, p. 24). NIH policy and resources about data sharing are available at http://grants.nih.gov/grants/policy/data_sharing/.
4. For two recent examples see "To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering" available at http://www.arl.org/bm~doc/digdatarpt.pdf and "Agenda for Developing E-Science in Research Libraries: ARL Joint Task Force on Library Support for E-Science Final Report and Recommendations" available at http://www.arl.org/bm~doc/ARL_EScience_final.pdf. The ARL and the Coalition for Networked Information will cosponsor the forum "Reinventing Science Librarianship: Models for the Future" in October 2008; Wendy Lougee, University Librarian (UMN) will convene and moderate several of the discussions. "The forum organizers hope to broaden the understanding of trends [including E-Science and Data Curation] in scientific research as well as support leadership in applying these trends in the development of new library roles" (¶ 1; http://www.arl.org/events/fallforum/forum08/index).
5. See http://www1.umn.edu/systemwide/strategic_positioning/summary.html.
6. See https://wiki.lib.umn.edu/Staff/Planning#FY09.
7. See https://wiki.lib.umn.edu/wupl/Staff.Planning/ULib_Compact07_StaffWeb_FINAL.doc.
8. The author was the cohort representative for the physical sciences and engineering.
9. As part of the cohort's observations of "emerging needs," the issues of data management and stewardship at the university level became apparent. These issues were later voiced by some UMN faculty during the "science assessment" study discussed next.
10. Description of the Data Services Librarian position courtesy of Amy West, Data Services Librarian, UMN Libraries.
11. The final report of this study can be found at http://purl.umn.edu/5546. Abbreviated summary reports, which explore the main themes of the study including that of data organization, management and preservation, are available at http://www.lib.umn.edu/about/scieval/documents.html. The "Science Assessment" study interviews UMN's scientific community in a manner similar to that of an earlier study of humanities and social sciences researchers at UMN, funded by the Andrew W. Mellon Foundation —" A Multi-Dimensional Framework for Academic Support" available at http://purl.umn.edu/5540.
12. A presentation on the data themes from the science assessment study occurred in May 2008, at the Center for Library Initiatives Conference, "Librarians and e-Science: Focusing Towards 2020" available at http://www.cic.uiuc.edu/programs/CenterForLibraryInitiatives/Archive/ConferencePresentation/Conference2008/e-ScienceSpeakerPresentations/delserone13May08.pdf.
13. See http://conservancy.umn.edu. The UDC is a DSpace instance.
14. For an example, browse the Minnesota Geological Survey's "Miscellaneous Maps" series, found at http://conservancy.umn.edu/handle/708, which includes GIS data files (in ArcInfo export format), but no supporting GIS applications (e.g., ArcView) are available.
15. See the spring 2008 Task Force Meeting of the Coalition for Networked Information available at http://www.cni.org/tfms/2008a.spring/abstracts/handouts/CNI_Assessing_Butler.pdf.
16. To view the slides from the RCA's presentation, including general sketches of each of the four research groups, see http://www.cni.org/tfms/2008a.spring/abstracts/presentations/cni-cyberinfrastructure=celeste.pdf.

References

Center for Library Initiatives Conference. (2008). Librarians and e-Science: Focusing Towards 2020. Retrieved October 31, 2008, from http://www.cic.uiuc.edu/programs/CenterForLibraryInitiatives/Archive/ConferencePresentation/Conference2008/e-ScienceSpeakerPresentations/delserone13May08.pdf

Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G. Fox & T. Hey (Eds.), *Grid computing: making the global infrastructure a reality* (pp. 809-824). New York: Wiley. Retrieved October 31, 2008, from http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf

Microsoft Research. (2004). Towards 2020 science. Retrieved October 31, 2008, from http://research.microsoft.com/towards2020science/downloads.htm

National Science Board. (2005). Long-lived digital data collections: enabling research and education in the 21st century. Retrieved October 31, 2008, from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf

NSF Cyberinfrastructure Council. (2006). *NSF's cyberinfrastructure vision for 21st century discovery* (version 7.1). Retrieved October 31, 2008, from http://www.nsf.gov/od/oci/ci-v7.pdf

University of Minnesota, Academic Programs. (2008, September 29). Reinventing reference. Retrieved October 31, 2008, from https://wiki.lib.umn.edu/AP/ReinventingReference

University of Minnesota, Academic Programs. (2008, October 13). Information literacy. Retrieved October 31, 2008, from https://wiki.lib.umn.edu/AP/InformationLiteracy

University of Minnesota, Academic Programs. (2008, October 15). Diversity Outreach Collaborative. Retrieved October 31, 2008, from https://wiki.lib.umn.edu/AP/DiversityOutreachCollaborative

University of Minnesota, E-Science and Data Services Collaborative. (2008, July 7). E-Science and Data Services Collaborative home page. Retrieved October 31, 2008, from https://wiki.lib.umn.edu/E-Science/HomePage

University of Minnesota Libraries. (2006). Compact for the University Libraries, 2006-2007. Retrieved October 31, 2008, from http://wiki.lib.umn.edu/wupl/Staff.Planning/ULib_Compact07_StaffWeb_FINAL.doc

University of Minnesota Libraries. (2007). Understanding research behaviors, information resources, and service needs of scientists and graduate students: A study by the University of Minnesota Libraries. Retrieved October 31, 2008, from http://purl.umn.edu/5546

University of Minnesota Research Cyberinfrastructure Alliance. (2008a). *Assessing research cyberinfrastructure needs at the University of Minnesota.* Retrieved October 31, 2008, from http://www.cni.org/tfms/2008a.spring/abstracts/handouts/CNI_Assessing_Butler.pdf

University of Minnesota Research Cyberinfrastructure Alliance. (2008b). *Assessing research cyberinfrastructure needs at the University of Minnesota.* Retrieved October 31, 2008, from http://www.cni.org/tfms/2008a.spring/abstracts/presentations/cni-cyberinfrastructure-celeste.pdf

University of Minnesota, Scholarly Communication. (2008, October 9). Scholarly communication home page. Retrieved October 31, 2008, from https://wiki.lib.umn.edu/ScholarlyCommunication/HomePage

---

Leslie M. Delserone is the agriculture librarian, University of Minnesota (UMN) Libraries, a member of the UMN Libraries' "E-Science and Data Services Collaborative," a former member of the Libraries' "University Digital Conservancy Working Group," and a current member of the IFPRI/UMN team developing *HarvestChoice*, an international project funded primarily by the Bill and Melinda Gates Foundation that "generates knowledge products to help guide strategic investments to improve the well-being of poor people in sub-Saharan Africa and South Asia through more productive and profitable farming" (http://www.harvestchoice.org/about/harvestchoice/ataglance.html). She earned her MA in Library and Information Science (University of Iowa) as part of the IMLS-funded "Program for University Librarians in the Sciences," after "prior lives" as an undergraduate academic advisor for the School of Biological Sciences, University of Nebraska–Lincoln and as a plant pathologist (MS, The Pennsylvania State University).