Jun 22nd, 10:30 AM - 11:30 AM

e-Science at the University of Minnesota: a collaborative approach

Lisa Johnston
University of Minnesota, ljohnsto@umn.edu

Cody Hanson
University of Minnesota, hans1794@umn.edu

“E-Science at the University of Minnesota: A collaborative approach.” (USA)

Lisa Johnston and Cody Hanson, University of Minnesota Libraries

Abstract

In 2008 the University of Minnesota Libraries formed the E-science and Data Services Collaborative (EDSC). The group was formed amid an environment of emerging initiatives related to e-science at the University, and was intended to leverage our existing expertise, such as our nationally recognized assessments of researcher behavior, to develop new capacity and engage with campus partners to support e-science and data services. We will report on the EDSC’s progress to date, including the following four areas of focus:

- A Data Stewardship Report assessing requirements for support of e-science and data services, determining gaps in our capacity, and seeking out opportunities to develop necessary expertise including data curation, data preservation, data policies and virtual organizations.
- A staff education program assessing the needs of libraries staff related to e-science and data services to establish a position description framework that includes E-scholarship: a potential new model for library liaison roles across campus that supports interdisciplinary and data intensive research.
- In coordination with the University’s Research Cyberinfrastructure Alliance (members include the Libraries, Office of Information Technology, Office of the Vice President for Research, and Minnesota Super Computing Institute), a two-phased plan for a Web Development project that defines our core services and areas of expertise in “data services” in the context of other campus services and initiatives.
- Increasing campus awareness of data management issues through the creation of a library Web site and skills-based workshop for faculty, students and researchers about data management best practices and university policies, including those that support open data initiatives.

INTRODUCTION

There has been a significant change in the way that research is done. Thanks to advancements in technology and the Web, scientific research in the digital age, or e-science, has become increasingly collaborative and carried out virtually with shared access to large data collections, distributed computing resources and, consequently, an increasing need for accountability at the research data level. In 2009, the National Academy of Sciences published its report, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age* (NAS, 2009), outlining the need for better data management practices and research transparency for the future sustainability of data in the digital age. The recommendations outlined in the report define the need for research data to be managed, shared, and retained for future use.

E-science, with its unfettered, new approaches to research problems, has also caused changes in thinking of the role that the library will play in this new research environment. A critical facet for research libraries is data stewardship, an activity that encompasses both preservation and curation of digital content. The "Agenda for Developing E-Science in Research Libraries" (Lougee, 2007), the Association of Research Libraries recommended the need and urgency for research libraries to take an active role in the education and policy surrounding e-science. To secure the future of library involvement with the curation and preservation of e-science data, the report signified a call-to-action prompting research libraries to

undertake similar methods of surveying the nature of e-science and cyberinfrastructure within their local domain and develop the skills needed for libraries to move into the emerging landscape.

At the University of Minnesota, the UMN Libraries have been actively engaged in understanding and supporting the evolving needs of researchers for the past several years. With a focus on the humanities, the libraries report "A Multi-Dimensional Framework for Academic Support" developed a model for assessing support for scholarship and research in the context of a large research campus (UMN Libraries, 2006). Then in 2006, with new funding from the Provost, the UMN Libraries recruited a cohort of three science librarians dedicated to supporting interdisciplinary research (Delserone, 2009). The cohort participated as focus group moderators in the library's study of university faculty and graduate students across scientific disciplines. In particular, the sciences assessment investigated the necessary elements of infrastructure that support the discovery, use and management of information sources and data, as well as uncovered the personal and often idiosyncratic repertoires and preferences of individual scientists (Marcus, 2007). During this time, the UMN Libraries began a Mellon funded pilot project focused on developing a virtual community for scholars in the interdisciplinary field of Bioethics. EthicShare (http://www.ethicshare.org) is a partnership with the University of Minnesota's Center for Bioethics, the University's Department of Computer Science and Engineering, and other institutions to allow for cross-institutional collaboration between researchers, providing a shared repository of citation and other bibliographic metadata (EthicShare, 2007).

Across the University of Minnesota campus were a number of developments related to e-science. Initiatives such as an agricultural initiative HarvestChoice (http://harvestchoice.org), the NSF DataNet grant proposal submitted jointly by the Minnesota Population Center and the UMN Libraries, the UMN Interdisciplinary Informatics Program, an increasing interest in GIS applications and scholarship, and the development of a program for the Clinical and Translational Science in the Academic Health Center (CTSA) have provided opportunity for library engagement. Finally, the formation of the campus Research Cyberinfrastructure Alliance (RCA) whose goal was to position the University to enable computationally intensive, interdisciplinary research for the 21st Century (RCA, 2008), provided the UMN Libraries a significant opportunity to articulate our particular expertise (current or aspirational) in developing this critical infrastructure.

E-SCIENCE AND DATA SERVICES COLLABORATIVE

Amid an environment of emerging initiatives related to e-science at the University, and responding to the national strategic priorities outlined by the ARL, in 2008 the UMN Libraries formed the E-science and Data Services Collaborative (EDSC). The goal of our group to develop new capacity and engage with campus partners to support e-science and data services (EDSC Wiki, 2008).

The EDSC was one of several ongoing collaborative efforts at the UMN Libraries. The collaborative working-group model brings together appointed staff with existing strengths from across the library system to provide leadership and guidance on far-reaching topics such as scholarly communications, reinventing reference services, information literacy, and diversity outreach. The focus of a collaborative is broad and often a moving target, therefore smaller project groups often set-out to accomplish specific tasks. In the EDSC, our ten-member team responded to the broad range of e-science and data service opportunities by forming five subgroups, four of which are discussed in detail here, including data stewardship and services, staff education, campus partnerships, and training and outreach. The fifth topic, GIS initiatives on campus, was primarily handled outside of our group with collaboration efforts between the UMN Libraries and the Geography department.

Data Stewardship

The range of potential activities implicated by an organizational commitment to the support of e-science is broad. The collaborative was tasked with assessing user-needs for e-science services, particularly related to data management and stewardship. To this end, our sub-group generated a report, “Data Stewardship Opportunities for the University of Minnesota Libraries,” in February, 2009.

The Data Stewardship report outlined six potential areas for library involvement. These areas are neither comprehensive nor mutually exclusive, and but demonstrated strengths of library staff and observed needs of the University community. The six opportunities were:

1. **Data Audits**
   By partnering with smaller departments or research groups on data audits, the UMN Libraries could gain a better understanding of the behavior of our researchers and of their complete data life-cycle. Our report recommended that the Libraries base these audits on JISC's Data Audit Framework, as implemented at the University of Edinburgh (JISC, 2009). The goal of these pilot audits would be to engage researchers with a hands-on process, to familiarize Libraries staff with the audit process and data life-cycle, and to create early successes that could be used to model and market larger-scale audits in the future.

   As of this paper, the data audits are underway with a joint-effort between the Office of Information Technology and the UMN Libraries. Once the audits are complete, OIT will experiment with the university's capacity to provide large-scale storage to individual researchers (potentially one terabyte of space) and also test the libraries data archiving capacity.

2. **Data Archiving**
   The UMN Libraries host a number of digital archives, including the University Digital Conservancy, our institutional DSpace repository built to archive text-based documents (UDC, 2010). Outside of the University, the UMN Libraries have a successful partnership in creating the HathiTrust repository (http://www.hathitrust.org/) to archive items digitized through Google's partnership with the Committee on Institutional Cooperation (CIC, 2007). Despite this experience, nothing in our organization's history has prepared us for the accumulated demand for long-term archiving of research data.

   Our Data Stewardship report concluded that the Libraries should actively seek cooperative partnerships on campus and beyond for purposes of creating archives for University research data. To the extent that the library will be able to provide data archiving, it will be with the help of university partners such as with the Office of Information Technology, the College of Liberal Arts, and the members of the RCA. In addition, the UMN Libraries should continue to explore and promote existing subject-based repositories for research data.

   Since the release of the Data Stewardship report, our institutional repository has begun accepting data sets on a case-by-case basis, such as GIS files. In 2010, we plan to move the repository to a Fedora-based system, and the requirements for the new platform will be created with data collection in mind.

3. **Mandate Compliance Consultation**
   The National Science Foundation and National Institutes of Health are two of the University's largest sources of external funding, and each has data management recommendations in place.

The Data Stewardship report determined that the UMN Libraries could take advantage of our knowledge of metadata standards and data archiving to provide guidance and support for researchers as they write grants and conduct research. Beyond this, the Libraries could provide expertise in author and researcher rights in publication, state data sharing and privacy laws, and Freedom of Information Act compliance.

4. **Development of Metadata Expertise**
   Libraries in general have extensive experience with bibliographic and archival metadata standards, and, in our case, a lesser knowledge and experience with domain-specific research metadata. The Data Stewardship report acknowledged that while we ought to continue to develop expertise in specific science and social science metadata schema, we would better position ourselves to support the University's emphasis on interdisciplinary research by encouraging the use of semantic, cross-domain standards such as RDF and OWL.

5. **Promotion of E-Science Services**
   Perhaps the quickest win for the UMN Libraries would be to explore methods of packaging and marketing the services and expertise we already have. Effectively promoting our services to researchers would require that we cut across departmental and divisional boundaries within our organization, and positioning our IT staff and infrastructure as public services. The Data Stewardship report likewise acknowledged that we must be mindful of those areas where campus partners may be better positioned to support researchers, and must partner or cede responsibility as appropriate.

   Following on this recommendation, our group created a informative web site that gathers information about e-science and data services provided by the libraries and across campus.

6. **Data Stewardship Support**
   The EDSC identified a number of research projects that are ongoing or in the planning stages which could benefit from library support. Examples include grant proposals by faculty that explicitly reference our repository or digitization efforts underway with the Antarctic Geospatial Information Center. These projects may serve as useful examples, or test cases, for a more comprehensive future data stewardship effort. While the UMN Libraries are justifiably reluctant to make commitments to new archival activities, the Data Stewardship report suggested that we could perhaps make an explicit statement to researchers of our willingness to hold materials on a temporary basis, until an enterprise solution is developed or another appropriate repository identified.

Staff Education

Library staff education, both within the group and broadly across the library units was a first priority of the EDSC. At the onset of the group forming, the EDSC group members participated in professional development opportunities, including the CIC Library E-Science conference (CIC, 2008) and the University of Illinois's Data Curation Education Program (2008). Additionally, a group wiki was established to announce information about the EDSC's activities (EDSC Wiki, 2008) and to gather relevant information in the early planning stages of the group to support our communication efforts. To better assess the needs of librarian staff related to e-science and data services the group organized on-site visits to each liaison department and held a staff-wide "journal club" focusing on e-science.

Our staff discussions produced comments consistent with what we've heard in the 2007 UMN Libraries' Science Assessment study. There was a perceived need for data services across academic disciplines which involved support with data storage, data sharing mechanisms, and tools and space for preservation of data files and software. The requests for data support are varied. Library units are contacted about data at different points in the data life-cycle. For example, liaisons in the social sciences and professional programs gets many requests to purchase data products, but almost no requests for help with locally produced data. University archives typically get requests for help with primary records and their associated software/format issues, on the other hand, the humanities librarians see researcher in need of tools that will in turn create and analyze data. In the physical sciences and engineering, data issues were centered on supporting locally held data sets that were not currently archived or in a sharable format.

Overall, liaisons expressed concerns about being adequately educated on data access and preservation issues before approaching faculty. Our visits suggested areas of focus for future library staff education, if not questions to ask ourselves. These included, clarifying the relationship of the EDSC to the institutional repository, developing a glossary of terms relating to data, creating an opportunity to explore in-depth examples of data life-cycles from several disciplines, and defining relevant intellectual property issues, especially open access mandates for research data. These questions fed the Training and Outreach group's web site development efforts where a data glossary, specific data examples, and mandate comparison chart were made available for staff.

Finally, the staff education efforts demonstrated a need to incorporating data management into the liaison position description framework. The liaison role was updated to include E-scholarship, an area of library engagement that supports interdisciplinary and data intensive research for all research across campus (Williams, 2009). The new role suggested examples for successful librarian engagement with staff as:

- seeking opportunities to collaborate with data producers and repository contributors to develop cost-effective and efficient strategies for managing data and information (which may include partnering in grants that require intense information and data management).
- recruiting institutional scholarly output, research data and other content for inclusion in the University Libraries' digital archiving initiatives.
- collaborating in the design, implementation, and maintenance of online tools and services that meet the needs of discipline/interdisciplinary research communities.
- developing knowledge of current practice and future directions in e-scholarship and help to identify gaps in existing support. (Draft Position Description Framework, 2009)

Campus Partnerships

The UMN Libraries have been a part of the University of Minnesota's Research Cyberinfrastructure Alliance (RCA), a group that also includes representatives from the Office of Information Technology, the Office of the Vice President for Research, the College of Liberal Arts, and the Minnesota Supercomputing Institute. The RCA is intended to coordinate University support for data-intensive research through cooperation on policy, infrastructure and services.

The EDSC was tapped to develop a web site for the RCA that would define the university members' core services and provide researchers with a single directory of resources to support their data-intensive work. We began by reviewing and updating a web content inventory of relevant existing content on RCA members' web sites, noting the name, description, and URL for each identified support service.

Recognizing that availability and cost would likely be paramount in the minds of the site's intended audience, EDSC members attempted to ascertain details for each service. Because of differing budget models, some relevant University services are free to researchers, while others are fee-based. Furthermore, these fees can vary (or disappear altogether) depending on a researcher's affiliation with a particular college or department on campus, and certain services may be available only to researchers affiliated with a particular college.

Our group also recognized that the terminology used to name and describe these services were full of jargon. Attempted to organize the web site using language that would be accessible to researchers, we categorized the services into three categories: Application Support, Data & Information Services, and Servers & Infrastructure.

The revised content inventory, proposed site categories, and a mock-up of the site were presented to a meeting of the RCA, where they were overwhelmingly approved. The EDSC group went on to build out the site. However, while the group was awaiting hosting and domain decisions, the RCA underwent a dramatic shift in organization and membership. It remains unclear what the future of the Libraries' E-Science initiatives are with respect to the RCA, and at the time of writing, the site developed by the EDSC group has yet to be deployed.

Training and Outreach

The training and outreach efforts of the EDSC grew organically over the course of the group's broader activities discuss above. The EDSC sub-group tasked with information gathering began by internally organizing reports and related information on the EDS wiki site. Also, the group established a social bookmarking account in Diigo (http://groups.diigo.com/group/escience) for all members to post, archive, and disseminate stories of interest.

Next, based on the information uncovered by the Data Stewardship group as well as a gap analysis of services provided by the RCA (2009), the information sup-group directed their efforts outward toward building awareness and education around e-science and data preservation issues on campus. Three primary outcomes of our efforts were: the creation of a library Web site, the development of a skills-based workshop for data management best practices, and a campus-wide presentation advocating the open data movement.

1. **EDSC Web Site "Managing Your Data"**
   Our goals for the library web site "Managing Your Data" (http://www2.lib.umn.edu/data/management) were to raise awareness and educate users on data management and sharing issues. The site presents information to both university researchers and library staff on the benefits of data management, collections and repositories for finding and sharing data, a glossary of data archiving and e-science terms, funding agency guidelines on data sharing, and copyright and ethical issues around research data. It also introduces researchers to common metadata standards, local storage and backup options, best practices for file naming and citation, and training opportunities in research computing around campus.

   Overall, the web site strives to inform the campus community about the libraries efforts to support e-science and present information on data management in order to facilitate best practices of data stewardship throughout the data life-cycle.

2. **Library Workshop "Introduction to Data Management"**
   The introductory data management workshop was developed by two EDSC members in the Physical Sciences and Engineering Library. Inspired by the workshop, "Data Management 101" (MIT Libraries, 2010) we set out to create a training opportunity for graduate students and faculty in the physical sciences and engineering disciplines. The workshop was created to reflect the university's services and established practices already presented on the "Managing Your Data" website.

   There was much interest generated by the workshop, in particular from the Office of the Vice President of Research, which saw this workshop as a potential fulfilment to the PI requirement of continuing education in the category of the ethical use of information. Due to their interest, our workshop has evolved to include a section on data use ethics with a case study of editing scientific images for publication. A next phase of the workshop will rework the focus toward researchers in the social sciences.

3. **University Presentation on "Open Data"**
   Most recently, in April 2010, the EDSC members who created the data management workshop presented at a University OIT Pecha Kucha event (http://www.oit.umn.edu/programs/20-by-20/) on the theme of "Open University." Our talk, "Open Data: Sharing Research for Greater Impact" was presented to a wide-audience on the benefits of sharing research data and how data management can make sharing easier. This outreach event was captured on video and the clip will be hosted on the Libraries' data web site.

CONCLUSION

Since the ESDC was established in 2008, priorities at the University and within the UMN Libraries have evolved. For example, there was a growing emphasis on E-scholarship, research-intensive support in all academic disciplines. Due to changes such as the Libraries "member" level involvement with digital humanities Project Bamboo (http://projectbamboo.org/) and the broadening focus for e-scholarship, the EDSC Group was disbanded in April 2010. The subsequent iteration, still in planning phases, will include the Digital Humanities as well as e-science and data issues.

The 2010-2011 UMN Libraries established the following E-scholarship goal: The Libraries will provide life-cycle management solutions for digital content through engagement in strategic partnerships, leveraging of Libraries' (and campus) assets, developing and sharing our expertise, and collaborating to develop essential infrastructure. The library's next steps toward E-scholarship include:

- Establish organizational structure to address program and infrastructure development.
- Clearly identify needs and opportunities through strategic market segmentation and analysis.
- Determine current capabilities and invest in the development of needed skills and capacity.
- Articulate functional roles including education and consultation and policy advocacy and advising.
- Implement data archiving and preservation services that position data for access and re-use; develop business plans firmly based on collaborations.

Although the EDSC met many of its initial goals and made great strides toward its established purpose, many questions remain. Despite the fact that our Data Stewardship Report gave us a more realistic sense

of current data archiving capacities and opportunities in e-science across campus, including social sciences and interdisciplinary science, we have yet to determine the limits of the institutional repository's role. Archiving unwieldy research data will require difficult decisions about what to discard, what to preserve, and for how long. We have yet to determine who is best equipped or empowered to make these decisions.

Collaborative-led staff discussions, brown bags, and wiki development began the process of build knowledge and capacity within the Libraries to support e-science and data services. However, it is clear that providing liaisons with the technical knowledge and support to make them comfortable discussing research data and digital archiving issues with faculty will require a robust ongoing education effort.

Finally, how does our work fit into the bigger picture campus-wide? The web development projects with the RCA and the "Managing Your Data" web site helped define the Libraries' core services and areas of expertise in the context of other campus services and initiatives. With outreach efforts like the Data Management workshop and Open University pecha kucha event, the group contributed to University discussions about interdisciplinary research and teaching and led the way toward developing a framework for educating campus about data policies including those that support open data initiatives. But is there a potential peril in successfully promoting these messages? We may be so lucky as to find that the Libraries' capacity for supporting e-science is outstripped by researcher demand.

There has been a fundamental transformation of how research is practiced in the digital age. Our experience at the University of Minnesota Libraries has shown us that stewardship of the evolving scholarly record will require ongoing education and organizational change. Whether e-science or e-scholarship, the underlying ideas of preservation and dissemination of research data will benefit from our involvement, a challenge we are committed to meet.

ACKNOWLEDGMENTS

The authors acknowledge the hard work and dedication of the University of Minnesota's E-Science and Data Services Collaborative, chaired by Kristi Jensen and Peter Kirlew, and whose members included: Leslie M. Delserone, Tony Fang, Gary Fouty, Cody Hanson, Amy Hribar, Lisa Johnston, Meghan Lafferty, Wayne Loftus, Jon Nichols, and Amy West. Thanks also to our collaborative sponsors: John Butler, Linda Watson and Karen Williams for their guidance and support.

REFERENCES

Cecily Marcus, et al. (2007). Understanding research behaviors, information resources, and service needs of scientists and graduate students: A study by the university of minnesota libraries. Accessed April 18, 2010, from http://www2.lib.umn.edu/about/scieval/documents.html

CIC Library Conference (2008). Librarians and E-science: Focusing Toward 20/20. Accessed on April 18, 2010 from http://www-s.cic.net/programs/centerforlibraryinitiatives/Archive/ConferencePresentation/Conference2008/home.shtml

CIC Web Site (2007). Committee on Institutional Cooperation Google Book Search Project - Introduction. Accessed on April 18, 2010 from http://www.cic.net/Home/Projects/Library/BookSearch/Introduction.aspx

Data Curation Education Program (2008). Graduate School of Library and Information Science Center for Informatics Research in Science and Scholarship at the University of Illinois at Urbana-Champaign. Accessed from http://cirss.lis.illinois.edu/CollMeta/dcep.htm.

EDSC Wiki (2008). About the Collaborative. Accessed April 18, 2010 from https://wiki.lib.umn.edu/E-Science/AboutTheCollaborative

EthicShare (2007) The Research Project. Accessed April 18, 2010, from http://www2.lib.umn.edu/about/ethicshare/index.html

JISC (Joint Information Systems Committee). (2009) Edinburgh Data Audit Implementation Project: Final Report. Project Report. Accessed from http://ie-repository.jisc.ac.uk/283/

Karen Williams (2009). Draft Position Description Framework. Unpublished document.

Karren Williams (2009) A Framework for Articulating New Library Roles. Research Library Issues: a bi-monthly report from ARL, CNI, and SPARC, 265, p3. Accessed April 18, 2010 from http://www.arl.org/bm~doc/rli-265-williams.pdf.

Leslie M. Delserone. (2008). At the watershed: Preparing for research data management and stewardship at the university of Minnesota libraries. Library Trends 57(2), 202-210. Accessed April 18, 2010, from Project MUSE database, http://muse.jhu.edu.floyd.lib.umn.edu/journals/library_trends/v057/57.2.delserone.html

MIT Libraries (2010). Managing Research Data 101. Presentation Slides accessed on April 18, 2010 from http://libraries.mit.edu/guides/subjects/data-management/Managing_Research_Data_101_IAP_2010.pdf

NAS (National Academy of Sciences) Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. (2009). Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Washington, D.C.: National Academy Press. Accessed April, 19 2010 from http://www.nap.edu/catalog/12615.html.

RCA (2008). Research Cyberinfrastructure Alliance. Accessed on April 18, 2010 from http://rca.umn.edu

RCA (2009). Draft Service Portfolio – Research Cyberinfrastructure Alliance. Internal Communication.

UDC (2010). The University Digital Conservancy. Accessed on April 18, 2010, from http://conservancy.umn.edu/aboutudc.jsp

University of Minnesota Libraries (2006). A Multi-Dimensional Framework for Academic Support: Final Report. Accessed on April 18, 2010 from http://conservancy.umn.edu/handle/5540.

Wendy Lougee, et al. (2007). Agenda for Developing E-Science in Research Libraries: Final Report and Recommendations to the Scholarly Communications Steering Committee, the Public Polices Affecting Research Libraries Steering Committee, and the Research, Teaching, and Learning Steering Committee. ARL. Accessed April 18, 2010, from http://www.arl.org/bm~doc/ARL_EScience_final.pdf