# Data Curation and Libraries: Short-Term Developments, Long-Term Prospects

**Anna Gold**
agold01@calpoly.edu
Associate Dean, Library
California Polytechnic State University, San Luis Obispo

**Abstract:**

This paper was prepared as background for a talk given at AGU 2009 on "Data & Libraries." It summarizes the developments and events from late 2006 through early 2010 that are shaping library roles in scientific data curation while underscoring the range, complexity, and varying granularity of systems, actions, and efforts involved. The main conclusions are: (1) leaders of major research libraries have committed their institutions to support data curation. (2) The library profession has demonstrated significant conceptual progress in characterizing and understanding data curation both in theory and in practice. (3) There has been progress since 2006 in legitimizing library roles in data curation through formal education and certification programs as well as by integrating data curation into established library services and systems. Certain questions remain unresolved: how will data taxonomies or ontology, schemas or data models and their databases fit into data curation practices? Librarians, however, can draw on a growing body of experience and the support of a community of practice as they contribute to data curation, while researchers and those who fund research can turn with growing confidence to libraries and librarians for data curation support.

Outline:

## 1. Introduction

The world is now approaching the 20th anniversary of the World Wide Web[1] and is fully engaged in the grand experiment it represents: a huge, complex, and constantly changing ecosystem of digital information and communication. The impact and significance of this digital ecosystem on the practice and communication of science – how we create and communicate research-based knowledge - is pervasive and in many ways revolutionary.

Libraries have long played critical roles within this ecosystem, in the pre-digital world of knowledge-making and knowledge-sharing. Libraries' role as custodians of "downstream" knowledge – the reports of research communicated in articles and books – is well established. Libraries also contribute to the knowledge lifecycle through less visible but crucial interventions "upstream," by advising researchers and teaching new scholars how to use the communicative apparatus of their field.

Today, not only libraries', but also many others' roles in the knowledge lifecycle must be reexamined in light of how that lifecycle is both exposed and transformed by the new digital ecosystem. A full exploration of the lifecycle of knowledge creation and use is beyond the scope of this paper, as is an exploration of concepts such as the data lifecycle or data curation lifecycle[2]. However, these concepts form an implicit backdrop for any discussion of the roles played by participants in data curation, from scientists, to funding agencies, and from publishers to libraries and archives.

Given this new digital ecosystem, and the changes it makes feasible in the production, use, and reuse of scientific data, what new roles can libraries play in the knowledge lifecycle? This is the question taken up in preliminary way in 2007 (Gold, 2007a and 2007b). Here we return to this question, beginning with a summary of developments from late 2006 through early 2010. Taken together these events and activities are shaping library roles in managing, preserving, and providing access to scientific data. While of primary interest to librarians, this summary can also inform scientists and others working with data management and data curation[3], both about recent developments, and their prospects over the long term.

It is apparent from a review of these developments that progress has been made, in the space of a few years, in the following areas:

---

[1] "On August 6, 1991, Berners-Lee posted a short summary of the World Wide Web project on the alt.hypertext newsgroup. This date also marked the debut of the Web as a publicly available service on the Internet." Retrieved on January 30, 2010, from http://en.wikipedia.org/wiki/History_of_the_World_Wide_Web

[2] For more on the knowledge lifecycle, see L. Lyon (2003) and C. Rusbridge (2005). Note also the white paper forthcoming in 2010 from OCLC Research and the U.K. Research Information Network (RIN) on support for research workflows (http://www.oclc.org/research/activities/support/).

[3] "Data curation" and "digital curation" may appear to be interchangeable terms. As used here, they refer to related but distinct concepts: "digital curation" is used for the curation of digital objects including compound digital objects; "data curation" for the curation of records or measurements of information ("data"). Those scientific measurements or records ("data") are further distinguished from the computer science meaning of "data" to refer to any type of digitally encoded information.

(1) Institutional leadership: There has been a steady and growing record of institutional actions by national leaders in library education and practice to secure a long-term role for libraries in acquiring and stewarding collections of scientific data;

(2) Conceptual progress: There has been significant progress in conceptualizing how library-managed institutional collections of scientific research can serve the needs of science within global educational, commercial, scientific, and technological infrastructures; and

(3) Legitimacy: There is an emerging sense of legitimacy regarding the social and technical roles that library professionals are being trained to play in support of scientific data curation, supported by an evolving formal curriculum.

### *1.1 Data curation, digital curation*

One of the challenges of talking about "data curation" is that the activities of curation are highly interconnected within a system of systems, including institutional, national, scientific, cultural, and social practices as well as economic and technological systems. Data curation is a nascent set of technologies and practices emerging in the context of this complex and rapidly evolving socio-technical ecosystem. By one definition, encompassing data in all formats,

> Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation and organization of these data for access and use over time. (Shreeves and Cragin, 2008, p. 93)

Qualifying "data curation" further as a type of *digital* curation adds a set of technical challenges shared with other types of digital objects:

> Digital curation, broadly interpreted, is about maintaining and adding value to a trusted body of digital information for current and future use. (DCC, "What is Digital Curation," 2007)

In its application to digital resources, including data,

> [C]uration embraces and goes beyond that of enhanced present-day re-use, and of archival responsibility, to embrace stewardship that adds value through the provision of context and linkage…in ways that ease re-use and promoting accountability and integration. (Rusbridge et al, 2005, p. 2)

At the same time, specifying that "digital curation" includes *data* curation causes the digital curation community to "extend our notions of curation" beyond static digital objects to resources that have "structure, volatility[4], and scale."[5] (DCC, "What is Digital Curation," 2007)

---

[4] "Volatility" or "data volatility" pertains to "the rate of change in the values of stored data over a period of time." See http://www.atis.org/glossary/definition.aspx?id=6970 (ATIS Telecom Glossary, 2007).

In recent years developments in the areas of both digital curation and data curation have come at a furious pace. It may be helpful to recap briefly some of the work that led up to this explosion of interest and activity.

Throughout the last decade of the 20th century, research funding was lavished on "digital libraries." In the U.S., major "DL" funding was distributed to mostly technological projects geared toward the development of digital collections of research materials, supported by specialized software and high performance hardware (see Griffin, 1998; and NSF/JISC, 2003).[6] By 2000, however, the opportunities and risks for science of a growing tsunami of digital data were being discussed with an increasing sense of urgency. Agencies that funded scientific research were particularly concerned about the fragility of their huge investments in the collection of scientific research digital data. In the U.S., these concerns set the stage for the National Science Foundation's Atkins report on cyberinfrastructure (Atkins, 2003) and culminated with the 2005 National Science Board's report on long-lived digital data collections (National Science Board, 2005).

Until this time, research libraries' concern with digital curation had been mostly focused on the challenges of preserving *digitized* objects that documented tangible library collections. This concern followed naturally from libraries' desire to preserve their own growing investment in these digital objects, including digitized images and texts. Libraries were also beginning to focus on the preservation of "born digital" editions of scholarly publications, especially journals. New arrangements among libraries and publishers were invented to assure the preservation of this content in the event, for example, of a publisher's failure (U.S.-based examples of these arrangements include NDIIPP, LOCKSS, CLOCKSS, Portico, the HATHI Trust, and PeDALS[7]). Outside of well established but relatively bounded library services in social science data and GIS, library roles in managing and preserving scientific data were relatively isolated and local practices (Gold, 2007b). The preservation of *data* had not emerged as a major concern of libraries.

This began to change as cross-institutional networked science, supported by national cyberinfrastructure, became an important feature of university-based research. University libraries began to consider that if more and more science was to be conducted, often collaboratively, using distributed computers and generating and processing huge amounts of data, it raised the questions of what kind of scientific record was being created and kept? and by whom? The idea began to take shape in research

---

[5] As thoughtful as these definitions are, the need for better definitions and consensus on the use of terms related to data is severe, and a recurrent theme to which I will return in the concluding section of this paper.

[6] An important critique of the NSF's Digital Library research program, articulated in a position paper by Carl Lagoze, was that it excluded web research – defined as research on the "interlinked knowledge network" that we know as "the web," whatever the current protocols and languages are that currently support that network. (Lagoze, 2003) The new focus on "cyberinfrastructure" to some degree remedies that instructive error.

[7] NDIIPP (http://www.digitalpreservation.gov/library/ ), LOCKSS (http://www.lockss.org/lockss/), CLOCKSS (http://www.clockss.org/clockss/ ), Portico (http://www.portico.org/), Hathi Trust (http://www.hathitrust.org /), PeDALS (http://pedalspreservation.org /).

libraries that the record of science in the future might include "libraries" of scientific research data. Would libraries' traditional mission of stewardship for the record of science extend to digital scientific data? If so, the challenge was difficult to overstate: it would mean going beyond *preserving* the digital bits (a problem libraries already knew was hard), to *curating* them. Curating would require making decisions to support selecting, sharing, and enabling new uses for those bits over time. But the payoff would be huge: a well-curated record of scientific data could itself become a new, vital and useful part of the process and practices of science, whether through data-mining and reuse, data-visualization, or other techniques and methods yet to be invented.

### *1.2 2007-2010: (trans)formative years*

By 2007 digital data curation roles for libraries and librarians had become a topic of new research, dialog, policy development, education, and debate, as indicated in Table 1 below. Libraries were participating in or organizing relevant conferences and workshops; librarians were producing and sharing new research in this area; and several new initiatives were launched to provide formal education opportunities for library professionals to prepare them to curate digital data.

In the U.S., the National Science Foundation (NSF) was the epicenter of influence. In the wake of the Atkins report and other major policy documents, the NSF began to strongly encourage libraries to play a role in data curation. Most notable was the NSF's broad call for participation by libraries when announcing its DataNet initiative, aimed at developing sustainable approaches to data curation. (NSF, 2007; Lee et al., 2009) The IMLS and the Mellon Foundation also encouraged these developments through significant program funding. International research, conferences, and policy development, notably in the U.K. and Australia, also exerted influences on the developing U.S. research library perspective on digital data curation.

Adding to these influences was increasing pressure from research funding bodies to require researchers to share their research data. Foremost among these in the U.S. was the NIH, with policies dating to 2003; other US funding bodies with this requirement include the Gordon and Betty Moore Foundation, in 2008, and the MacArthur Foundation, also in 2008.[8]

The following table (Table 1) briefly identifies major professional milestones that, from the end of 2006 through early 2010, have contributed to shaping the future of libraries in data curation.

[8] A global listing of data archiving policies of funding agencies can be found in SHERPA / Juliet (http://www.sherpa.ac.uk/juliet/).

Table 1: Libraries and Digital Data Curation:
Major Milestones in Education, Policy, Research, and Services, 2006 - 2010

| YEAR | AREA | MILESTONE DESCRIPTION |
|---|---|---|
| 2006 | Education | One of the nation's premiere graduate programs in librarianship, at UIUC, obtains an IMLS grant to develop its Data Curation Education Program, DCEP, October 2006. Five students enroll. [http://cirss.lis.illinois.edu/pdf/DCEP_Annual_Report_Year_1.pdf] |
| 2007 | Education | At ASIST annual meeting in November 2007, a session is held, "Identifying Best Practices and Skills for Workforce Development in Data Curation." [http://www.asis.org/Conferences/AM07/panels/41.html] Summary at: [http://databits.lternet.edu/node/78] |
| 2007 | Education | At the University of Illinois, Urbana Champaign, the Graduate School of Library and Information Science (GSLIS) program teaches a course, "Foundations of Data Curation" for the first time in fall 2007. [http://www.isrl.illinois.edu/~wjohn/] and [http://www.ideals.illinois.edu/bitstream/handle/2142/9716/prepescienceinfospec.ppt?sequence=2] |
| 2007 | Policy | The Association of Research Libraries (ARL) establishes the ARL Joint Task Force on Library Support for E–Science, charged to inform membership about E-Science, develop relationships with key stakeholders, recommend approaches to curating digital data, and engage in developing new roles and skills for library information professionals. [http://www.arl.org/rtl/escience/escicharge.shtml] |
| 2007 | Research | The National Science Foundation (NSF) publishes in January a report of the Sept-Oct 2006 NSF workshop, "History and Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures," urging that infrastructure requires the development of social institutions and practices and calls for partnerships with organizations that have "substantial existing expertise in areas complementary to scientific research." [http://www.si.umich.edu/InfrastructureWorkshop/ ] |
| 2007 | Research | The Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF-SDPA) is funded by NSF and the Mellon Foundation, in partnership with the Library of Congress, the UK's JISC, CLIR, and NARA. Multiple library community leaders are named to the Task Force, which issues its preliminary report at the end of 2008. [http://brtf.sdsc.edu/ ] |
| 2007 | Research | UIUC and Purdue, with support of Institute for Museum and Library Services (IMLS) funding, launch the Curation Profiles Project (2007-2009), to study differences between scientific domains and institutional cultures re. data curation problems and needs. [http://www.datacurationprofiles.org/] |
| 2007 | Research | NSF publishes its proposal for DataNet on September 28, 2007, envisioning "new types of organizations [that] will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise." Five awards totaling $100M are anticipated in two sequential years (FY09 and FY10), with awards paid over five years. [http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf07601] |
| 2007 | Research | NSF and the Mellon Foundation hold a Workshop on Scientific and Scholarly Workflow Cyberinfrastructure, Baltimore, Maryland October 3-5, 2007. A technical report from the workshop is published in 2007 While most attendees are scientists, there are key library leaders present. [https://spaces.internet2.edu/display/SciSchWorkflow] |
| 2007 | Research | NSF holds an informational meeting for potential applicants for DataNet funding, November 2007:. [http://www.nsf.gov/news/news_summ.jsp?cntn_id=110392] |

| | | |
|---|---|---|
| 2007 | Research | The Digital Curation Center (DCC), together with CNI, NSF, and IMLS, holds the 3rd International Digital Curation Conference in Washington, DC, December 11-13, 2007: "Curating our Digital Scientific Heritage: a Global Collaborative Challenge". [http://www.dcc.ac.uk/events/dcc-2007/ ] |
| 2007 | Research | JISC and the Mellon Foundation hold Workshop on Sharing and Curating Research Data, Washington, D.C., December 14, 2007. [http://infteam.jiscinvolve.org/files/2008/05/datacurationwshop20071214.pdf] |
| 2008 | Education | Syracuse University's School of Information Studies receives NSF CI-TEAM funding in March, 2008 to create prototype certification program for masters students as "Cyberinfrastructure Facilitators" [CI-Facilitators: Information Architects across the STEM Disciplines] [http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0753372] |
| 2008 | Education | The Digital Curation Centre (DCC) organizes an inaugural workshop, "Developing an International Curation and Preservation Training and Education Roadmap," in Washington DC, May 27-28, 2008:. Reported in D-Lib, March/April 2009. [http://www.dlib.org/dlib/march09/hank/03hank.html ] |
| 2008 | Education | UIUC's GSLIS program holds its first Summer Institute on Scientific Data Curation, June, 2008. [http://www.lis.illinois.edu/programs/cpd/DC_Inst/] |
| 2008 | Education | ARL/CNI hold their Fall Forum in Arlington Virginia October 16-17, 2008, on "Reinventing Science Librarianship," Arlington, Virginia. A summary of the proceedings is published in February 2009. [http://www.arl.org/events/fallforum/forum08/index.shtml] |
| 2008 | Education | UIUC (GSLIS) enrolls ten new students into their Specialization in Data Curation, September 2008. [http://www.lis.illinois.edu/programs/ms/data_curation.html] |
| 2008 | Education | Syracuse University School of Information Studies enrolls first five students in its CI-Facilitators program, September 2008. [http://ischool.syr.edu/media/documents/2008/12/HomepageWeb_Fall08.pdf] |
| 2008 | Education | Joint Conference on Digital Libraries (JCDL) annual meeting includes workshop, "Education for Digital Stewardship: Librarians, Archivists or Curators?" in Pittsburgh, PA, November 2008. [http://www.ils.unc.edu/jcdl2008/] |
| 2008 | Education | Second meeting of the International Data Curation Education Action (IDEA) Working Group meeting, December, 2008. Reported in March 2009. [http://www.dlib.org/dlib/march09/hank/03hank.html] |
| 2008 | Research | Metadata for Scientific Datasets (MeS) Workshop, held at the Dublin Core 2008 (DC-2008) Conference in Berlin, Germany, September 25, 2008. One outcome is resolution to create a DCMI Science and Metadata Community (SAM). A follow up meeting is scheduled for October 2009. [http://ils.unc.edu/mrc/sci_metadata/] |
| 2008 | Research | The U.S. National Research Council (NRC) forms a new Board on Research Data and Information (BRDI) in October 2008. Its mission is to help improve the management, policy, and use of digital data and information for science and the broader society. Board members include several information scientists and research library leaders. [http://sites.nationalacademies.org/PGA/brdi/index.htm] |
| 2008 | Research | The DCC with CNI holds the 4th International Digital Curation Conference in Edinburgh, December 1-3, 2008,"Radical Sharing: Transforming Science?" [http://www.dcc.ac.uk/events/dcc-2008/] |
| 2008 | Research | A Coalition for Networked Information (CNI) Task Force holds a group meeting in December 2008 to discuss the need for an international data registry service. [http://www.cni.org/tfms/2008b.fall/] |
| 2008 | Research | The Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF-SDPA) issues its interim report in December 2008. [http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf] |
| 2008 | Research | The NSF's National Science Board announces two major DataNet awards in December 2008 - one to DataONE and the other the Data Conservancy (led by the Johns Hopkins University Libraries), both to begin summer 2009. |

| | | |
|---|---|---|
| | | [http://www.cendi.gov/presentations/CENDI_03-10-09_Spengler_NSF_DataNet.pdf] |
| 2008 | Services | The Cornell University Libraries (CUL) Data Working Group publishes its final report in May, 2008, "Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library [http://ecommons.library.cornell.edu/bitstream/1813/10903/1/DaWG_WP_final.pdf ] |
| 2008 | Services | MIT Libraries launch a data management web site for MIT faculty, researchers, and others, August 2008. [http://libraries.mit.edu/guides/subjects/data-management/] |
| 2008 | Services | The Association of Research Libraries' E-Science Task Force publishes its report on E-Science, including E-Science Talking Points for ARL Deans & Directors, October 2008. [http://www.arl.org/rtl/escience/] |
| 2008 | Services | HATHI Trust launches, October 2008, committing to digital preservation of electronic texts. [http://www.hathitrust.org/] |
| 2009 | Research Services Policy | Following a workshop session on Metadata for Scientific Datasets at DC-2008, a new DCMI Community is established in February 2009, on metadata for scientific datasets - DCMI Science and Metadata Community (SAM). [http://ils.unc.edu/mrc/sci_metadata/] |
| 2009 | Education | UIUC's GSLIS holds its second Summer Institute on Data Curation in May 2009, this time in the Humanities. [http://cirss.lis.illinois.edu/CollMeta/dcep/SummerInstituteHumanities.htm] Blogged at: [http://cmsmcq.com/mib/?p=553] Slides by Dorothea Salo at: [http://www.slideshare.net/cavlec/digital-preservation-and-institutional-repositories] |
| 2009 | Education | ACRL's Science and Technology Section organizes a panel June, 2009 at ALA Annual meeting: "Big Science, Little Science, E-Science: The Science Librarian's Role in the Conversation." Background bibliography prepared by Denise Bennett (July 2009). [https://www.ideals.uiuc.edu/handle/2142/13145] [http://wikis.ala.org/acrl/index.php/STS_2009_Program] and [http://connect.ala.org/node/71879]; Selected session posters are published at: [http://www.ala.org/ala/mgrps/divs/acrl/about/sections/sts/conferences/posters09.cfm] |
| 2009 | Education | The Computer Science Roundtable Special Libraries Association (SLA) hosts a panel June 2009 at the annual SLA meeting: "Data Curation and Special Libraries: Education, Trends, and Developments." [http://units.sla.org/division/dpam/pam-bulletin/vol37/no1/computer.htm] |
| 2009 | Policy | The US National Science and Technology Council issues their Report of the Interagency Working Group on Digital Data to the Committee on Science of the NSTC, "Harnessing the Power of Digital Data for Science and Society," January 2009. [http://www.nitrd.gov/About/Harnessing_Power.aspx] |
| 2009 | Policy | The National Academies Board on Research Data and Information (BRDI) issues their report, "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age," January 2009. [http://www.nap.edu/catalog.php?record_id=12615] |
| 2009 | Research | DigCCurr 2009 (conference) is held April 1-3, 2009 in Chapel Hill, NC, on "Digital Curation Practice, Promise and Prospects." [http://www.ils.unc.edu/digccurr2009/] |
| 2009 | Research | The CNI Spring Task Force Meeting on April 6, 2009 includes on the Data Conservancy and DataONE projects (both anticipating funding under NSF's DataNet call) in Minneapolis, MN. [http://www.cni.org/tfms/2009a.spring/plenary.html] |
| 2009 | Research | An Open Repositories conference is held May, 2009 at Georgia Tech, funded in part through NSF's DataNet, for its promise to " lead to new and maturing partnerships that support digital repository advancements, increasing repositories' value as an essential element of cyberinfrastructure for research and education." [https://or09.library.gatech.edu/] |

| | | |
|---|---|---|
| 2009 | Research | As part of DataNet funding, NSF awards $299,688 in September 2009 to fund a feasibility study to create an open access National Science Foundation (NSF) publication repository. Johns Hopkins University's Sheridan Library leads, together with the Council on Library and Information Resources (CLIR) and the University of Michigan (UM). [http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0948134] |
| 2009 | Research, Education, Services | Organized by librarians at the University of Massachussetts, Amherst, NERCOMP hosts a regional (New England) event in October 2009, "Science Librarians in an e-Science World," featuring speakers from Cornell, Yale, MIT and the University of Massachusetts. [http://www.nercomp.org/events/event_single.aspx?id=5839 ] |
| 2009 | Research | Official launch in October 2009 of year one of first two NSF DataNet awards: DataONE and the Data Conservancy. Major library partners of DataONE are the California Digital Library; and of the Data Conservancy, JHU Libraries, Portico, UCLA, and UIUC library researchers. (Lee et al., 2009) |
| 2009 | Research | EDUCAUSE Annual Conference in November 2009 includes presentations on DataONE and Data Conservancy. [http://educause.mediasite.com/mediasite/SilverlightPlayer/Default.aspx?peid=663de84527a44b979bea7eeb6d715bf0 |
| 2009 | Research | ASIS&T annual meeting in November 2009 includes panels on DataNet and data curation. [http://www.asis.org/Conferences/AM09/panels/41.html] [http://www.asis.org/Conferences/AM09/open-proceedings/panels/27.xml] |
| 2009 | Research | IMLS awards funds to the University of North Carolina's School of Information and Library Science, with partners in the IMLS and the U.K.'s Digital Curation Centre. The proposal, "Closing the Digital Curation Gap (CDCG), is to establish baseline practices for the storage, maintenance, and preservation of digital data [http://www.imls.gov/news/2009/112009c.shtm] |
| 2009 | Services | Following input from the 2008 Research Libraries Group (RLG) Partnership meeting, members of the RLG Research Information Management Interest Group formed a working group to "define and advance" the issue of the role of libraries in data curation. [http://www.oclc.org/research/activities/datacuration/default.htm] |
| 2010 | Education | The Canadian Association of Research Libraries releases Research Data: Unseen Opportunities, in January 2010, as an "awareness toolkit" "to enable research library directors to raise awareness of the issues of data management with administrators and researchers on campus." [http://www.carl-abrc.ca/about/working_groups/pdf/data_mgt_toolkit.pdf] |
| 2010 | Research | BRTF final report published, "Sustainable economics for a digital planet: Ensuring long-term access to digital information," February, 2010. [http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf] |
| 2010 | Research | NSF announcement of three additional major awards in the DataNet program is anticipated in fall 2010. (Spengler, personal communication February 16, 2010) |
| 2010 | Research | A special summit on Research Data Access and Preservation, co-sponsored by CNI, is scheduled for April 9-10 in conjunction with the ASIS&T Information Architecture Summit. Phoenix, AZ. [http://www.asis.org/Conferences/IA10/ResearchDataAccessSummit2010.html] |
| 2010 | Education | Third Summer Institute on Data Curation at UIUC, May 2010. [http://listproc.ucdavis.edu/archives/iamslic/log1003/0025.html] |
| 2010 | Research | International Association of Scientific and Technological University Libraries (IATUL) conference, The Evolving World of e-Science: Impact and Implications for Science and Technology Libraries, at Purdue University and Chicago, June 20-24, 2010. [http://blogs.lib.purdue.edu/iatul2010/] |
| 2010 | Research | Digital Curation Conference, "Participation & Practice: growing the Curation Community through the data decade," December 6-8, 2010 in Chicago, hosted jointly by DCC and GSLIS at UIUC, in partnership with CNI. [http://www.dcc.ac.uk/events/dcc-2010/index.php] |

## 2. Short term developments in data curation roles for libraries

At the beginning of the second decade of the 21st century, a relatively small number of research libraries have staff who are directly and actively involved in digital data curation. Most research scientists are unaware that libraries are capable of playing such a role. So is it reasonable to question whether digital data curation will have a place in library practice outside of a few research libraries and library research and education programs?

Table 1 suggests that the answer to that question is "yes." The sustained support of data curation-related activities sponsored by the ARL, CNI, and professional organizations (ACM/IEEE, ASIS&T, ACRL), together with the developments in graduate library education programs supporting data curation, indicate that digital data curation has lodged deeply in the heart of the research library community.

The library systems and library researchers of several major research universities, including Johns Hopkins University, Cornell University, the University of Tennessee, the University of California (through the California Digital Library as well as several campuses), and the University of New Mexico, have lead roles as research and development partners in projects funded through NSF's DataNet program.

Also during this period, several graduate library and information science programs have piloted new training initiatives to develop the skills and knowledge needed to support the curation of digital data.

While research and education are likely to remain a primary focus of data curation activity by academic libraries in the immediate future, the number of professional library publications and events related to data curation suggests that a longer-term place for libraries in digital data curation is emerging.

Alongside what could be characterized as "elite" data curation activities represented by the work of the DCC in the U.K., or by those U.S. libraries with lead roles in DataNet projects, libraries are demonstrating a growing interest in providing local data curation services on their campuses. In addition, the emerging relationship between library-run institutional repositories and national or global networks of data repositories suggests that the practice of digital data curation will make rapid progress – as a library practice – over the next decade, (Baker and Yarmey, 2009; Witt, 2008)

In the next few years, these developments suggest that roles for librarians in digital data curation will fall into one or more of three tiers:

- *National infrastructure:* A small number of research libraries, working with government bodies, professional organizations, and industry, will have a large role in helping to formulate national digital data curation strategies, including economic models to support curation over the long term.

- *Campus infrastructure:* A larger number of libraries and librarians will actively support the development of campus-based data curation services. In developing these services, they will be able to draw on a growing set of resources created by research library leadership (e.g. ARL) and leaders in the campus technology

community (e.g. CNI, EDUCAUSE).

- *Professional development and education:* Graduate programs in library and information science are developing to support professional roles in data curation. Library leaders have opened national dialog and invested in both formal education and continuing professional development. Also notable is the broad level of engagement of individual librarians, who are actively working to increase their own data literacy and awareness, and equipping themselves to provide educational and consultative services related to data management and curation to their students and faculty. They are organizing and attending data-related workshops and conferences, conducting research into faculty data curation needs, and teaching basic data management skills to their students or faculty. While individually these efforts are on a modest scale, the grassroots engagement of librarians in these issues is one of the most exciting developments of the last several years, with the potential to shape both the future data practices of scientists and the practice of librarianship.

Below, specific examples and details further illustrate these three levels of engagement by libraries.

2.1 Libraries as strategic partners in national data curation strategies

In the US, most digital data curation research has been funded by three organizations: the National Science Foundation (NSF), the Institute for Museum and Library Services (IMLS), and the Mellon Foundation. These influential funding agencies have all placed an emphasis on strategies of partnership and collaboration, and on the economic sustainability of digital data curation as a major challenge.

A primary example is the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF-SDPA, funded in part by both the NSF and the Mellon Foundation) Reporting in 2008, Task Force member Fran Berman stated that "researchers need help with things Librarians are good at," and identified those "things" as:

- Developing reliable management, preservation, and use environments;
- [Assuring] proper curation and annotation;
- Navigating policy, regulation, intellectual property;
- [Facilitating] collaboration (partnership to share resources, create economies of scale, etc.); and
- [Assuring] sustainability.

The focus of the BRTF-SDPA on developing real world economic models for meeting digital preservation needs illustrates this shift in attention from a focus on the technical challenges of digital data preservation. This shift has also been signaled by early stages of public discussion of the DataNet program, where historic strengths of libraries and archives were noted as keys to sustainable data curation: reliability, expertise in resource sharing, policy development, annotation and selection, and institutional commitment to sustaining access over long periods of time. (NSF, 2007b)

It is clear that even if research libraries have these areas of expertise, and have enjoyed relatively stable, long-term funding streams that support traditional library services, most libraries are unlikely to be in a position to curate major collections of digital data

themselves. Research library funding can be fragile, especially in difficult economic times. The stresses of inflation in the cost of research journals, compounded by the recent recession, have stripped many libraries of any excess capacity; and the significant costs of retooling staff and infrastructure to curate a digital record of unprecedented complexity and diversity are sobering.

Thus major data curation funding agencies do not anticipate that libraries will necessarily own, curate, and manage major digital data repositories. Rather, they hope that libraries will play a facilitating role in establishing collaborative networks of organizations that will be capable of executing this responsibility.

The NSF's 2007 call for proposals in its DataNet program was entitled "Sustainable Digital Data Preservation and Access Network Partners." The solicitation aimed to create a set of exemplar national research infrastructure organizations that would "integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise." (NSF, 2007a)

Clifford Lynch described the proposal as a "classic, large-scale NSF initiative" proposing to make five awards over five years, totaling $100M. (Choudhury and Lynch, 2009) But the goal of DataNet is not only to do "capacity-building" for large-scale data curation, but also to develop curation models that will scale and extend across a wide range of disciplines, and that will, at the end of their funding cycle, result in usable systems and become economically sustainable, independent of NSF funding. (Spengler, 2009) One speculation is therefore that, unlike the mega-initiatives of the NSF Digital Library program (Griffin 1998), this NSF initiative promises both to engage, and to apply directly to the interests of smaller, non-research libraries.

In its FAQ for potential DataNet applicants, the NSF clarified that they had no "one-size-fits-all" approach to the balance they expected funded projects to demonstrate between domain science expertise and preservation and access/infrastructure:

> We believe in user/use-centered design. We also believe that librarians, archivists, and computer/ computational/information scientists are unlikely to build excellent infrastructure for science and/or engineering without deep engagement with the intended users. In that sense, domain scientists should be full partners in the process. (NSF, 2008)

For research libraries interested in data curation, the level of library involvement in the first two DataNet awards announced in December 2008 was impressive.

One of the awards was made to the Data Conservancy project, led by Saheed Choudhury, Associate Dean of the Sheridan Library at Johns Hopkins. Partners in the Data Conservancy proposal include library researchers at UIUC and UCLA as well as library-based software initiatives (DuraSpace, Portico), the British Library, and a number of scholarly publishers.

Officially launched in October 2009, the Data Conservancy proposes to connect and adapt existing systems and standards. It plans to develop these through user-centered design and research, and sustain them through a portfolio of "funding streams." Forming an interoperating network, these systems and standards will be managed through a coordinated governance structure. (Choudhury, 2009) Technically, the Data

Conservancy expects to be built using principles of modularity; with layers of storage; a common conceptual framework ("observations") for managing data across a wide variety of domains; and making use of the proposed OAI-ORE (Object Reuse and Exchange) mapping standard for compound objects, including data.

The second major project funded by NSF through the DataNet program is DataONE (Data Observation Network for Earth). It too includes significant roles for research libraries and leading library researchers. Led by William (Bill) Michener, Professor and Director of E-Science Initiatives for University Libraries at the University of New Mexico, the DataONE leadership team includes Suzie Allard and Carol Tenopir, both of the University of Tennessee (Knoxville) School of Information Sciences, and Patricia Cruse of the California Digital Library. As with the Data Conservancy, DataONE's goal is to create a distributed, sustainable cyberinfrastructure, in this case focused on "well-described and easily discovered Earth observational data." DataONE leaders describe the project as constructing a "virtual data center," using integrated finding tools for data using a variety of metadata standards[9]. They have announced plans to integrate "downstream" tools such as experiment workflow sharing (My Experiment) in order to enable sharing and replication of workflows and results. DataONE will also emphasize education and outreach to scientists, "creating an informatics-literate workforce through innovative outreach and training efforts (e.g., best-practice videos, podcasts, on-line certificate programs, downloadable best practice guides and exemplars of data management plans)." DataONE is inviting libraries and organizations as well as individuals to become "member nodes," with access to software and instructional materials.

Both the DataONE and the Data Conservancy teams have emphasized that the infrastructure they are building is not an end in itself, but a means to help researchers address the grand, interdisciplinary research challenges that face society.

Although the first two DataNet projects are still at an early stage the project goals shared by project leaders underscore that these awards anticipate long-term roles for research libraries in data curation, and that these will include:

1. Supporting <u>interoperability of metadata</u> for scientific observations to support <u>cross-domain and cross-community</u> search and discovery;

2. Developing <u>metadata standards for complex research data records</u> that relate recorded observations to published analyses as well as to various related entities and descriptors;[10]

3. <u>Consulting</u> with individual researchers and research groups on <u>best practices for data management</u>;

---

[9] FGDC-Biological Data Profile, EML, GCMD, Z39.50, Darwin Core, Dublin Core, and ISO 19115; see early prototype at http://mercdev3.ornl.gov/dataone/

[10] Including Open Archives Initiative Object Reuse and Exchange (OAI-ORE). See Pepe et al., 2009; also http://www.openarchives.org/ore/. Another metadata research initiatives to be developed through the DataONE DataNet initiative is Dryad, which like ORE proposes a standard for representing the relationship between different outputs of the research process, such as a dataset, a working paper, and a published journal article (Greenberg, 2009).

4. Contributing as data scientists to ongoing research teams by advising on and developing team practices and policies to support both immediate and future data curation, reflecting domain practices and needs;

5. Developing data use cases that will inform design goals and principles for planned data curation infrastructure; and

6. Collecting digital data. In an age where so much data is at risk, libraries can contribute to the goals of data curation by "gathering as much data as you can," even if the data is small in size and its future use unclear. (Choudhury, 2009)

As additional DataNet projects are launched, the next five years should see a multiplication of instances of these roles in practice, along with the emergence of communities of practice supporting these roles.

At the same time, one of the challenges of implementing strategies of partnership and collaboration is the need to build awareness and understanding across different sectors. It is important, for example, to acknowledge the tremendous amount of research data that is already deposited in and managed by national archives and government data centers. Such centers may be found in national libraries or international organizations (e.g. the World Data Center for Climate[11]). In the U.S., examples include not only the National Archives and Records Administration (NARA) but also scientific data centers such as NOAA's National Climatic Data Center[12], or the National Snow and Ice Data Center[13]. The partnership of libraries with such centers through the Data Conservancy may provide opportunities for exactly the kind of boundary-crossing collaboration that is essential to sustainable systems for data curation.

2.2 Campus-based data curation.

Like funding agencies, university campuses have a tremendous investment in the production of scientific data[14]. Campus IT infrastructures are coming under pressure to meet data management and preservation services. Major organizational challenges of providing such services include the wide range of data types, scientific domains represented, the scale of scientific digital data collection, and the expertise required on any one campus to provide such services. Even apart from technical challenges (network, bandwidth, storage, preservation), campus staff may be asked to:

- Help faculty access cyberinfrastructure services locally (and, when necessary, globally);
- Assist faculty in managing their data—including observational data, the construction of research and reference collections, or data from analysis or

[11] http://www.mad.zmaw.de/wdc-for-climate/

[12] NCDC, http://www.ncdc.noaa.gov/oa/ncdc.html

[13]NSIDC, http://nsidc.org/

[14] Campus IT organizations are also responsible of course for digital administrative data centers and services as well as digital data associated with humanities and creative work (e.g. CAD files of architects, digital recordings of music or other performances, or historical data).

simulation—and preparing this data for handoff to the appropriate data repositories and curators at the appropriate time; and
- Aid faculty in parallelizing computations or organizing data for reuse, mining, and mashups.

"Probably the greatest challenge of cyberinfrastructure at the campus level will be the design and staffing of the organizations that will work with the faculty," writes Clifford Lynch in 2008. To meet campus needs, existing staff will require "more expertise in disciplinary data, standards, and tools and perhaps also with more capability for consulting on software, data, and information design.... Given these requirements for scale, one final set of questions concerns staff: Where will the necessary staff come from? How will they be trained? What educational qualifications and background will they have, and what academic programs will produce them?" (Lynch, 2008[15])

Some research libraries, among them libraries of Cornell, Purdue, the Massachusetts Institute of Technology (MIT), and the University of Minnesota, have taken the initiative to convene campus-wide E-Science initiatives or centers, or initiate data curation partnerships with domain researchers, computer scientists, and campus IT.[16] The following brief descriptions provide a representative, though by no means a comprehensive, picture of the range and type of these activities.

*Cornell*: In 2005, supported by NSF funding for exploratory research, librarians at Cornell began pursuing data management initiatives for data in ecology, linguistics, and (power) blackout research. With additional NSF funding, Cornell later developed a "data staging repository," DataStaR, to support collaboration and data sharing throughout the research process, as well as serving as a model for academic libraries to provide a "transitory" curation environment:

> "The model leverages the ability of a researcher's local institution to provide accessible support and services related to research data, early in the research process, and serves to promote the deposition of data in domain-specific repositories, thus making data available to the larger research community."

The model offers the best of both worlds, preserving the advantages of local knowledge while also promoting the transmission of data to repositories better suited for long-term curation and preservation.[17]

In a presentation at the spring 2007 meeting of the Coalition for Networked Information (CNI), Cornell staff described the value libraries brought to campus data initiatives:

---

[15] See also the EDUCAUSE Campus Cyberinfrastructure Working Group, http://www.educause.edu/CCI.

[16] In addition to campus initiatives there are multi-campus cyberinfrastructure initiatives such as TIP, a two-year multi-campus initiative to develop cyberinfrastructure for research data for several universities and research centers in North Carolina. University librarians as well as the CIO's and provosts for the three partner campuses were signatories to this plan. See press release of October 26, 2009, retrieved January 24, 2010 from http://www.renci.org/news/releases/data-initiative.

[17] http://datastar.mannlib.cornell.edu/

> ...the Library brings to the table credibility as a neutral and a competent information broker, in the game for the long haul. As data stewardship and full lifecycle information management become essential to research competitiveness and even mandated by federal funding agencies, research libraries have an important leadership role to play. Libraries will need new skills and above all new partners, within and beyond our own universities. The tasks will challenge both our own and our partners' traditional thinking, but in many ways the future of the research library will depend on acquiring, preserving, and delivering the data and knowledge essential to the research enterprise today. (McCue and Corson-Rikert, 2007)

Another data initiative at Cornell University Libraries was the formation in 2006 of a Data Working Group (DaWG) in order to exchange information about data curation and to recommend "strategic opportunities for CUL to engage in the area of data curation." [18]In 2009 the Working Group released an extended white paper, summarizing their findings and recommendations. In addition to providing an outstanding review of current education, issues, and research in data curation, the white paper provided a narrative audit of major institutional data activities at Cornell.

The Working Group recommended that Cornell University Libraries partner with other university data producers and data centers; and also that they provide a variety of data services to the Cornell University community, including:

- Assisting with formulating researchers' data management plans;
- Collecting and providing best practices information for data management;
- Educating researchers on intellectual property issues related to data;
- Offering informed referrals for services not provided by the libraries; and
- Working with other university partners to formulate institutional data archiving policies.

In addition, the Working Group recommended that the Libraries form a Data Curation Executive Group; and that they establish a leadership position with responsibility for data curation and e-scholarship (Steinhart et al., 2008).

*Purdue:* At the end of 2006, the Purdue University Libraries established a "Distributed Data Curation Center" (D2C2) (http://d2c2.lib.purdue.edu/) in partnership with faculty in computer science and other disciplines (Mullins, 2007). D2C2 was the centerpiece of an ambitious initiative that drew Purdue subject librarians into active roles supporting the data curation needs of campus researchers. For example, Purdue Libraries staff developed experience with conducting systematic data interviews (Witt 2009; Garritano and Carlson, 2009). In 2008, together with researchers at UIUC's library and information science (LIS) graduate school, Purdue Libraries received funding from the IMLS to conduct a two-year Curation Profiles Project, focused on learning about the variables that affect researchers' willingness to share their data. (Witt et al., 2009) In 2009 Purdue University Libraries also launched "e-Data," an institutional data

---

[18] While Cornell's was among the first such groups, there were and are a growing number groups formed for similar purposes. They include groups at Ohio State University Library; at Georgia Tech; at the University of North Carolina, at UIUC, and at the University of California, Irvine (STS-L listerv, February 6, 2010; Melissa Cragin, personal communication, February 6, 2010).

repository, using a different platform than those selected for other types of digital assets.[19]

*Massachusetts Institute of Technology:* An informal Data Initiatives Group (DIG) began meeting in 2006 to conduct data interviews with MIT researchers and to share information on developments in data services across many science and engineering domains, including chemistry and bioinformatics. By 2008 the group was managing an extensive data services portal on the MIT Libraries' web site. (MIT Libraries, 2008) This approach to campus data service development leverages and transforms the traditional subject liaison role of librarians into "data liaisons." (Gabridge, 2009) The research group within the MIT Libraries has also played a major role in library digital preservation research, as a developer of DSpace and SIMILE. The MIT Libraries research program is now exploring the role of institutional repositories for managing research data (Smith, 2009). In 2009 the MIT Libraries also participated in the "How much information" (HMI) project[20] funded by a research consortium led by the University of California, San Diego. Working with an MIT faculty member and an MIT graduate student, the MIT team developed and published case studies of data creation and use across six science and engineering disciplines. (Madnick et al., 2009)

*University of Minnesota:* The Libraries of the University of Minnesota are leading a campus-wide Research Cyberinfrastructure Alliance that is charged to conduct research on campus infrastructure needs and to explore possible data service models. In addition the Libraries are conducting case studies of researchers' data practices (Lougee, 2008); and have begun a staff education and reorganization effort that in 2008 established an E-Science and Data Services Collaborative[21]. The Collaborative's goals include:

- Building knowledge and capacity within the Libraries to support E-Science and data services, including knowledge of scientists' research needs, funding agency data stewardship requirements, and metadata standards for data in the sciences;
- Defining core services and areas of expertise in "data services" in the context of other campus services and initiatives;
- Defining a potential new model for library liaison roles across campus that supports interdisciplinary science (including relevant social sciences); and
- Contributing to University discussions about interdisciplinary research and teaching and developing a framework for educating the campus about data policies, including those that support open data initiatives.

In addition to the libraries of the universities featured here, there are many other university libraries that are engaged in preparing themselves for a future in digital data curation through research, staff education, and organizational realignment. Among these are the University of Massachusetts at Amherst (Schmidt and Reznik-Zellen, 2009), and the University of Virginia (Sallans et al., 2009). Institutional activities like

[19] http://www4.lib.purdue.edu/lcris/edata/

[20] http://hmi.ucsd.edu/

[21] EDSC, https://wiki.lib.umn.edu/E-Science/

these are everywhere supplemented by research initiated by individual librarians.[22] While some pieces of data curation infrastructure may exist in university libraries (e.g. institutional repository software, metadata services, specialists in domain informatics), applying this infrastructure to digital data curation unavoidably requires an investment in learning. This investment is essential if libraries are to retool their skills and reframe their roles, both within their campuses, and in relationship to scientific working practices and scientific digital data collection.

One of the opportunities in local campus communities is for librarians and other professionals (including project information managers) to forge communities of practice with benefits to both. Just as librarians can play a role in bridging the "last mile" to researchers as consultants on data management resources (Gabridge, 2009), they can also create alliances with data management and information management professionals associated with campus research groups. Standards-making at this level may be a particularly productive and critical as an area of collaboration, especially in light of research findings that underscore that:

> The diversity of data types, working methods, curation practices and content skills found even within specialized domains means that [data curation] requirements should be defined at this or even a finer-grained level, such as the research group. (Key Perspectives, 2010, p. 3)

2.3 Education and professional development

In September 2008 a technical report to JISC offered the following assessment of the "current practice and future needs" for skills, roles, and career structure of data scientists and curators:

> The library and information science community should have an important role to play in the data science arena, particularly in delivering awareness and understanding of data issues and the importance of good data science and data curation. There are generic data handling and management skills that are native to librarians and can be taught as part of the basic research skills training in an institution. After all, the fundamentals of data science can be taught and subject expertise can be acquired over time. There are also other roles that libraries can play here. We suggest that three of the most relevant ways in which the library community might influence developments are:
>
> - Training researchers to be more data-aware[23]

[22] For example, at Cal Poly San Luis Obispo a science librarian and the librarian responsible for Cal Poly's institutional repository are partnering with a statistics faculty member to conduct a pilot survey of scientists' data curation practices and needs (J. Scaramozzino and M. Ramirez, personal communication, January 2010).

[23] "Data awareness" is an element of a broadened conception of "information literacy" that embraces awareness, interpretation, and responsible reuse of scientific data.
Baker and Millerand point out that the NSF criterion of "broader impacts" for the evaluation of scientific research could be extended to encompass training and outreach not only to lay communities but also to scientific communities, in order to foster the types of information literacies that scientists will need to make effective use of data across a variety of scientific disciplines and practices. (Baker and Millerand, forthcoming)

- Adopting a data archiving and preservation role [and]
- The training and supply of data librarians.

(Swan and Brown, 2008)

A significant development in building the case for library roles in data curation and for creating the capacity to carry out those roles to was the formation by the Association of Research Libraries (ARL) in 2007 of a Joint Task Force on Library Support for E–Science, charged to inform membership about E-Science, develop relationships with key stakeholders, recommend approaches to curation of digital data, and engage in developing new roles and skills for library information professionals. The Task Force provided an opportunity for library leaders to share thinking and expertise on E-Science and digital data and potential roles for research libraries in both. One of the charges to the Task Force was to create educational opportunities for research librarians and leadership in this area; and to develop talking points for library directors to use in discussing the role of libraries in E-Science and digital data curation with their campus leadership. The Task Force completed its charge in 2008, holding a national workshop on the changing roles of science librarians in light of E-Science and digital data curation; and published both its final report and a set of talking points in the same year.

In the U.S., several other library professional organizations are devoting increasing resources and attention to digital data curation. The American Society for Information Science and Technology (ASIS&T) annually hosts posters, panels, and research presentations on data curation. In 2010 ASIS&T will be sponsoring a new data curation event, a summit on digital data curation to be held in conjunction with their annual Information Architecture (IA) Summit. The Science and Technology Section of the Association of College and Research Libraries (ACRL) Division of the American Library Association hosts regular presentations on digital data curation, as does the Special Libraries Association. In 2010, the annual meeting International Association of Scientific and Technological University Libraries hosted by Purdue University (IATUL) will focus on libraries and E-Science.

In the U.K., the Digital Curation Centre (DCC) is an essential resource for librarians with an interest in data curation, hosting regular conferences and workshops and hosting an open access international journal on data curation.[24]

More universities now offer formal education aimed variously at developing specialists trained (and certified) in digital data curation, introducing practicing professionals to data curation, or bringing practitioners together for advanced training.

In 2006, the Graduate School of Library and Information Science (GSLIS) at the University of Illinois, Champaign-Urbana (UIUC), obtained an IMLS grant to develop a Data Curation Education Program (DCEP). By fall 2007, the program enrolled five students, and it began offering a course on the "Foundations of Data Curation" (Cragin et al., 2007). Ten new students were enrolled by fall 2008 into their Specialization in Data Curation. In addition, the UIUC GSLIS program began offering a practitioner-oriented data curation education program in 2008, when they held a first Summer Institute on Scientific Data Curation. The following summer the institute focused on data curation in

---

[24] IJDC, http://www.ijdc.net/index.php/ijdc

the humanities, and another data curation institute is planned for summer 2010 (Cragin et al., 2009).

Also in 2006 the University of North Carolina (UNC) launched the first phase of DigCCurr (pronounced "dij-seeker") with funding from the IMLS and the National Archives and Records Administration. DigCCurr is charged to "develop an openly accessible, graduate-level curricular framework, course modules, and experiential and enrichment components and exemplars necessary to prepare students to work in the 21st century environment of trusted digital and data repositories." While the focus of DigCCurr extends beyond scientific data to include cultural artifacts and records, cultural heritage assets, and teaching materials, it also encompasses curation of research data. The project has sponsored national symposia in 2007 and 2009 to "bring the issues of digital curation and this curriculum to the broader library, archives, and museum communities as well as the public." A second phase, DigCCurr II, will develop "an international, doctoral-level curriculum and educational network in the management and preservation of digital materials across their life cycle." Among its activities are one-week professional institutes aimed at bringing practitioners together (June 2009, also planned in May 2010 and January 2011).

Other new and developing graduate education programs include:

- *Syracuse University:* Syracuse received funding in 2008 to develop a program of internships and training to support the development of "cyberinfrastructure faciltators," defined as "one who aids the research of topical experts with cyberinfrastructure." Among the professional credentials the program will offer are a 2-year masters program, and certification as a "CI-facilitator."

- *University of Michigan's School of Information*: Plans were announced to develop a new course by 2010 in science/social science data curation, as part of new Specialization in Information Preservation (Yakel et al., 2009).

- *Australian National Data Service (ANDS):* An initiative began in 2008 to build capabilities to support data curation through developing a curriculum and certification process, working with major stakeholders such as CSIRO and GeoScience Australia (Burton and Henty, 2009).

Whether through participation in professional conferences and workshops or through formal education, practicing librarians can now:

- Develop an understanding of local data management practices, using as a model a variety of interviewing questions and approaches shared by the DCC, Purdue, Cornell, MIT, and others;

- Develop an understanding of both the technical capacity of local institutional repositories (IR's) to manage digital data, and the policy and procedural issues regarding the "right role" of IR's as long-term or transitional digital data repositories;

- Develop sufficient understanding of data curation best practices and the relevant policies of funding agencies, to advise and refer students and researchers at local campuses and research centers;

- Develop and share an appreciation of the value of "small science" data, in light of national DataNet infrastructure building, and as reflected in the "web of repositories" perspective (Baker and Yarmey, 2009) and Heidorn's description of the "long tail" of digital (and dark) data (Heidorn, 2009).

In addition to supporting current and future researchers, librarians play a vital role in helping new scholars and students gain information skills and fluency, as well as awareness of information and communication practices that are essential for effective work in a domain. In areas such as GIS, social science data, and bioinformatics, these roles are relatively well established (e.g. Hunt, 2004). At an American Libraries Association (ALA) panel in 2009, Melissa Cragin described a growing role for academic librarians in advancing "data literacy" (Cragin, 2009) As secondary users and mediators of curated data, librarians help students and researchers: search for and retrieve curated data; select data for reuse, by assessing its quality as well as its "fit" to the information need; manipulate data using a variety of technologies and tools; cite and attribute data accurately; and provide consultation and training to potential data users.

Such educational activities by librarians blend information literacy with science literacy education. A related initiative in this area is the Science Data Literacy (SDL) project at Syracuse[25] funded by the National Science Foundation. The SDL project researchers have planned and offered a new undergraduate course open to students from multiple science disciplines, "Science Data Management," and are sharing the course syllabus on the web[26]. Taught in several modules the first offers an overview of data fundamentals (including forms, scales, types, data structures and models, and data formats); the second uses case studies to teach about data aggregations at the research, resource, and reference levels; and the third introduces methods for evaluating data quality and using data in different communities of practice.

## 3. Long term prospects

### 3.1 Data curation community of practice:

Taken as a whole, the developments described above illustrate that a data curation community of practice is emerging that includes wide representation from the library and information professions. To a large degree this community has emerged in and around national centers such as the U.K.'s Digital Curation Center, as well as around national research funding programs (as with DataNet in the U.S.). There is also a growing grassroots community of practice in professional library and information science organizations and education programs.

In addition, one of the most interesting developments in digital data curation is the emphasis placed by all parties on developing a community of practice. While centers and funding projects play key roles, there is also widespread support for "bottom-up" alliances that support data curation. An example is the DuraSpace Data Curation Solution Community, which its organizers describe as "based on the theory that higher

[25] http://sdl.syr.edu/

[26] http://sdl.syr.edu/?page_id=15

levels of order will emerge from complex systems under the right conditions," and adding that:

> Data curation should support new forms of research and learning across disciplines ranging from the sciences to the humanities. Requirements must be gathered from both professional and citizen researchers and learners who may also participate with data curation infrastructure development.[27]

Another community space, the Digital Curation Exchange, is hosted at the University of North Carolina (home of DigCCurr), with similar goals for the digital curation community. Their goal is to serve as a " 'town center' for the practitioners, researchers, educators, and students of digital curation," with discussion forums, a place to exchange educational modules, and other resources of interest.[28] A third wiki for sharing information on digital data curation is hosted at the University of Oregon.[29]

The significance of the development of communities of practice for data curation goes beyond the immediate benefits of sharing information and problem-solving. It also signals the role played in data curation by networks as an essential strategy for making sense across boundaries, while permitting change to take place within an unbounded, nonhierarchical, and loosely coupled system (Baker and Millerand, forthcoming)

3.2 Services, not systems:

In the U.S., NSF-funded digital library research shifted over the course of major funding projects from being technology- and collection-centered, to focusing more on users, services and service layers, through programs such as the National Science Digital Library (NSDL). (NSF/JISC, 2003) Learning from this experience, NSF's emerging digital data curation program emphasis is on building services, not systems. The clear message is that collecting and keeping data is an intermediate goal, with curating for reuse and cross-disciplinary use the underlying and more fundamental goal (Choudhury and Lynch, 2009)

For libraries, this shift is significant in that it mirrors a broader transformation in the view of research libraries not as primarily repositories for information but as active agents who provide both opportunities and infrastructure to support the exchange of ideas and knowledge by research communities. One illustration of this new perspective is the statement by the University of California's California Digital Library (CDL) that it has re-articulated the mission of its Digital Preservation Program mission:

> ...in terms of digital curation, rather than preservation; encouraging a programmatic, rather than a project-oriented approach to curation activities; and a renewed emphasis on services, rather than systems. ... The Program is pursuing a path towards a new curation environment based on the principle of devolving curation function to *a set of small, simple, loosely coupled services*.... (Abrams et al., 2009)

---

[27] http://fedora-commons.org/confluence/display/FCCWG/Data+Curation

[28] http://digitalcurationexchange.org/

[29] http://libwiki.uoregon.edu/display/DigCur/Data+Curation+Home

The match between libraries' deep knowledge base and expertise in the areas of metadata specification and development, user education and outreach, and user needs analysis, is a good indicator that libraries are well-matched to undertake long term roles in supporting data curation.

### 3.3 The future of institutional repositories

Institutional repositories (IRs), created and managed by libraries in order to provide open access to the research conducted at a university, seem poised to play a useful role in data curation. Given the wide variation in data needs and practices across domains of research, and the acceptance of national and global data centers within many domains, the role of IR's in data curation has to date been modest. In contrast to institutional repositories, cross-institutional repositories managed by and for scientists in a particular domain[30] have in some cases been far more successful. However, a "both / and" approach to IR's and domain repositories is emerging, with growing acceptance that campus IR's can play an important role as "feeders" of data to discipline repositories. Among those articulating this vision is Cornell University Library, with their DataSTaR initiative (described above); as well as Baker and Yarmey (2009), whose "web of repositories" model articulates how the important initial advantages of local data management can be integrated with a vision of long-term curation in domain-managed collections. (Steinhart et al., 2009)

Similarly, the Australian National Data Service (ANDS) is creating an Australian Research Data Commons to unite the data repositories of institutions (libraries and universities) and discipline or domain organizations (ANDS Technical Working Group, 2007).

Taken together, these developments demonstrate the three major achievements for libraries in data curation in recent years noted earlier: first, the emergence of strong institutional leadership at all levels within the library and information communities for a library role and voice in a data curation community of practice; second, progress within that community in conceptualizing the problem of data curation in terms of overarching service goals rather than technology; and third, establishing the legitimacy of library roles in data curation through formal education and training as well as by integrating data curation into existing library services such as institutional repositories or research consultative services.

### 3.4 Interoperability and boundary-crossing partnerships

Another conclusion can also be drawn from the evidence of data curation activity in the last several years. This is that libraries' participation in cyberinfrastructure for data curation will draw them into a set of practices and relationships with other sectors and professions in new ways.

It may be that the scale, complexity, and distribution of the challenges of data curation present libraries with the opportunity to practice strategies they can apply to other information management and service challenges. For example, early lessons of work on

---

[30] The physics preprint arXiv, though now managed by Cornell University Library, is the canonical example of a domain repository for text [http://arxiv.org/]; among the many discipline-managed data archives is the Protein Data Bank [http://www.pdb.org/]

data curation and cyberinfrastructure emphasize the importance of interoperability, of user-centered design, and of boundary-crossing partnerships. Also, the scale of the data curation challenge is so large that it will have to be sustained by a portfolio of funding streams, and may be managed, in part, by a coordinated governance structure. It may be helpful to anticipate that as a multi-level structure through a federation, "defined as the act of uniting multiple states or sites where each retains control over its own internal affairs." (Baker and Millerand, forthcoming) Looser, sometimes ad hoc networking and collaborations may play important roles, with networks extending across national and public-private boundaries. Even libraries that are not directly participating in developing data curation services at this time will find these developments relevant in planning their own futures. Rather than asking the question, "what role will libraries have in cyberinfrastructure?" a better question now may be, "how will libraries' practice of cyberinfrastructure transform their future?"

## 4. Some risks and problems

While the opportunities of data curation are great, this emerging field also poses some risks to those who want to be part of addressing its challenges. Perhaps the greatest risks are associated with creating solutions before understanding fundamental problems. Among the problems that need further understanding are:

*1. Data sharing:* An assumption of any type of data curation activity is the desirability, at least over time, of sharing data beyond its author or authors. Despite a common commitment to sharing scientific findings, the readiness or willingness of researchers to share data varies widely. Reasons for this are often characterized in terms of researchers' desires to retain intellectual property or intellectual ownership, and are labeled "data withholding." However Baker and Millerand point out that there are also "scientifically salient concerns" with sharing data, including "resistance to propagating ill-described data to an audience unfamiliar with the field's data handling issues." (Baker and Millerand, forthcoming). Addressing such underlying concerns with sharing is essential to any meaningful data curation strategy.

*2. Coevolution:* Coevolution[31] is a good description of the complicated interactions between various systems in relation to digital scientific data. Every human, technological, and economic system involved in data curation, from mark-up languages to the social movement for open science, is part of a digital information ecology in which every part evolves in a field of mutual influences. Coevolution is a much larger context than intentional networking and planned collaboration. It is important that coevolution be factored into data curation goals, for example by taking care that solutions developed today are not dependent on unfounded assumptions about stability in other sectors.

*3. Metadata and schemas:* A great challenge of data curation is ensuring that data, once preserved, remains meaningful either within the same research area or ideally across areas or even across domains. The expectation that descriptions of data (metadata) can do this work for every kind of data is problematic, both because metadata encodes

[31] The term "coevolution" appeared in the 1960's, initially in biology but eventually to describe mutual influences in technical and cultural change, or the "joint evolution of two or more systems that interact with each other" (The American Heritage New Dictionary of Cultural Literacy, Third Edition).

implicit formats that may not be shared across research areas; and also because any type or level of data description limits the amount and type of information available about the data, and excludes information that might be relevant to another researcher. In addition, the raw data is given meaning in part by the schema (or model) representing relationships between the data. The role of data schema in enabling data reuse merits further study (see Gray, 2009).

*4. Controlled vocabularies and taxonomies:* As individuals and groups from widely varying domains and professions work together to address the complexities of data curation, it is important that they understand as clearly as possible what they mean when they use terms for data-related concepts and categories. Clarifying what is meant by "data," "curation," "data acquisition," "data production," and the like may seem unnecessary, but this clarification is important ongoing work.[32] Building up shared taxonomies that reveal how data curation concepts are related, whether in hierarchical or nonhierarchical ways, is also valuable work. If this work is done in ways that includes reference to concrete particulars, it is more likely that all parties will gain a more accurate but also nuanced understanding of their own and others' use of common terms (Pennington, 2008). Of course, multiple perspectives and definitions may necessarily coexist for some time, and meaningful distinctions are worth preserving and honoring.

*5. Ontologies.* If scientific data curation is both for, and about, science, then the ontologies represented by scientific data cannot be ignored. How we conceive and process information about the world shapes the data we collect, and informs our efforts to express our views using instruments that include the machines and computer programs that process our data. How will data be connected across domains that are based on different ontologies? In a gesture echoing the pragmatic field of library science, the Data Conservancy may place its bets on the "observation" (either human or by proxy) as a universal atomic entity for data curators (Choudhury and Lynch, 2009), just as the entities resulting from "authorship" form the bedrock enabling cross-domain research in conventional libraries. But the practical question, at large, is managing in swift order petabytes of (often heterogeneous) data. In doing this, the file format might be a more practical universal atomic entity for both automation and observation. Other approaches to connecting data across domains and ontologies include the "Concept Web Alliance," described by Van de Sompel and Lagoze (2009).

## 5. Conclusion

The last several years have been marked by a steady and growing record of institutional actions by library graduate schools and national library leaders to secure a long-term role for libraries in acquiring and stewarding collections of scientific data. As a result of these actions as well as local developments and widespread professional discussion and education, the library profession has made considerable progress in conceptualizing how library professionals and library-managed institutional collections of scientific research can serve the needs of science within global educational, commercial, scientific, and technological infrastructures. As further illustration of this progress, a formal curriculum for training and education is emerging, and positions for professional librarians are being advertised to support data curation programs and services.

Over the second decade of the 21st century, library investments in data curation will surely vary, but each investment will represent an addition to a more general cultural change in libraries. This change is characterized by a greater priority assigned to

stewarding locally created research (Lewis, 2007), by a shift in emphasis from information literacy to fluency in the processes of science and culture; and by a growing commitment to facilitating the reusability of information and data. Future librarians will draw on a growing body of experience and the support of a community of practice as they play valued roles in data curation. Researchers can turn with growing confidence to their librarians, knowing that librarians can, and want, to play roles in supporting the mutual goal of data curation.

The big bang of the web is still expanding in all directions with increasingly complex and fruitful implications for libraries as organizations whose ultimate mission is not to keep, but to share knowledge. In years ahead, digital information and data will continue to expand in scale, diversity, and potential uses, though the artifacts that convey and enable those uses will remain vulnerable for some time to come. In understanding and then responding to these challenges, the world's libraries are bcoming part of a broadly collaborative and long-lived process of stabilizing and cultivating relationships between digital data and human meaning.

## 6. References and further reading (* denotes recommended readings)

Abrams, S., P. Cruse, and J. Kunze (2008). Preservation is not a place. International Journal of Digital Curation, 4(1), 8-21. Retrieved January 15, 2010, from http://www.ijdc.net/index.php/ijdc/article/view/98

Allinson, J. (2008). Describing scholarly works with Dublin Core: A functional approach. Library Trends 57 (2), Fall 2008, pp. 221-243. Retrieved January 15, 2010, from http://muse.jhu.edu/journals/library_trends/v057/57.2.allinson.html

Altman, M. and G. King (2007). A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine 13 (3/4), March/April 2007. Retrieved November 15, 2009, from http://www.dlib.org/dlib/march07/altman/03altman.html

Anderson, C. (2004). The Long Tail. Wired, 12 ( 10), October 2004.

ANDS Technical Working Group (2007). Towards the Australian Data Commons: A proposal for an Australian National Data Service, October 2007. Australian Government, Department of Education, Science, and Training. Retrieved January 23, 2010, from http://www.pfc.org.au/pub/Main/Data/TowardstheAustralianDataCommons.pdf,

Atkins, D. (2003). Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Retrieved December 15, 2009, from http://www.nsf.gov/od/oci/reports/toc.jsp

Baker, K. and F. Millerand (forthcoming, 2010). Infrastructuring ecology: Challenges in achieving data sharing. To be published in: Collaboration in the new life sciences, ed. J. Parker et al., Ashgate.

*Baker, K., and L. Yarmey. (2009). Data stewardship: Environmental data curation and a web-of-repositories. International Journal of Digital Curation, 4(2). Retrieved November 22, 2009, from http://www.ijdc.net/index.php/ijdc/article/view/115.

*Beagrie, N. Blog. Retrieved April 4, 2010 from http://blog.beagrie.com/

*Beagrie, N. (2006). Digital curation for science, digital libraries, and individuals. International Journal of Digital Curation, 1 (1). Retrieved April 4, 2010, from http://www.ijdc.net/ijdc/article/view/6/5

Beagrie, N. (2007). E-infrastructure strategy for research. Final report from the OSI [Office for Science and Innovation] Preservation and Curation Working Group. Retrieved April 4, 2010, from http://www.nesc.ac.uk/documents/OSI/preservation.pdf

Beagrie, N., J. Chruszcz, and B. Lavoie (2008). Keeping research data safe: A cost model and guidance for UK Universities. Retrieved April 4, 2010, from http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx

Berman, F. (2008). Got data? A guide to data preservation in the information age. Communications of the ACM, 51(12). Retrieved February 6, 2010 from http://en.scientificcommons.org/52469420

Berman, F. (2008). Research and data. Presented at ARL/CNI Fall Forum in Arlington Virginia October 16-17, 2008, "Reinventing Science Librarianship," Arlington, Virginia. Retrieved January 24, 2010 from http://brtf.sdsc.edu/pubs/ARL-CNI08.pdf

Burton, A. and M. Henty, (2009). Building Australia's eResearch capability: The challenge of data management. Presented at EDUCAUSE Australia, May 3-6 2009. Retrieved January 23, 2010, from http://www.caudit.edu.au/educauseaustralasia09/assets/papers/monday/Margaret-Henty.pdf.

Choudhury, G. S. (2008). Case study in data curation at Johns Hopkins University. Library Trends 57 (2), Fall 2008, p. 211-220. Retrieved December 10, 2009, from http://muse.jhu.edu/journals/library_trends/v057/57.2.choudhury.html

Choudhury, G. S., and C. Lynch (2009). Initiatives from the NSF's DataNet program: DataONE and the Data Conservancy. Retrieved November 12, 2009, from http://www.educause.edu/E09+Hybrid/EDUCAUSE2009FacetoFaceConferen/InitiativesfromtheNSFsDataNetP/175757 .

Corson-Rikert, J. and J. McCue (2007). Engaging and connecting faculty: Research, discovery, access, re-use, and archiving, CNI, Spring 2007. Retrieved February 6, 2010, from http://www.cni.org/tfms/2007a.spring/abstracts/handouts/CNI_Engaging_McCue.pdf

Cragin, M. H. (2009). Panel presentation for Big Science, Little Science, E-Science: The Science Librarian's Role in the Conversation, Science and Technology Section, American Library Association, July 13, 2009. . Retrieved February 6, 2010 from http://www.ala.org/ala/mgrps/divs/acrl/about/sections/sts/conferences/Cragin_ALA_STS_Chica.pdf

Cragin, M. H., Palmer, C. L., Heidorn, P. B., and Smith, L. C. (2007). An educational program on data curation. Poster at American Library Assocation, Science and Technology Section, summer 2007. Retrieved January 30, 2010, from http://www.ideals.uiuc.edu/handle/2142/3493

Cragin, M. H., Smith, L. C., Palmer, C. L., and Heidorn, P. B. (2009). Extending the data curation curriculum to practicing LIS professionals. DigCCurr 2009 proceedings, 92-93.

DataStaR (n.d.). Cornell University Libraries. Retrieved January 30, 2010, from http://datastar.mannlib.cornell.edu/

Delserone, L. M. (2008). At the watershed: Preparing for research data management and stewardship at the University of Minnesota Libraries. Library Trends 57 (2), Fall 2008, pp. 202-210. Retrieved January 30, 2010, from http://muse.jhu.edu/journals/library_trends/v057/57.2.delserone.html

Digital Curation Centre (2007). About the DCC: What is digital curation? (2007). Retrieved November 11, 2009, from http://www.dcc.ac.uk/about/what/

Edwards, P., S. Jackson, G. Bowker, and C. Knobel (2007). "Understanding infrastructure: Dynamics, tensions, and design. Report of a workshop on 'History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructure.'" Retrieved December 15, 2009, from http://www.si.umich.edu/InfrastructureWorkshop/

Foster, S., Jelinkova, K., & Russel, C. (n.d.). Considerations for campus cyberinfrastructure data management policy and procedure development. EDUCAUSE. Retrieved November 12, 2009, from http://www.educause.edu/Resources/ConsiderationsforCampusCyberin/173213 .

*Gabridge, T. (2009). The last mile: The liaison role in curating science and engineering research data. Research Library Issues, 265 (Aug. 2009). Retrieved February 6, 2010, from http://www.arl.org/bm~doc/rli-265-gabridge.pdf

Gandel, P., Stanton, J., Lankes, R. D., Cogburn, D., Liddy, E., and Oakleaf, M. (n.d.). More people, not just more stuff: Developing a new vision for research cyberinfrastructure. EDUCAUSE.

Retrieved November 14, 2009, from
http://www.educause.edu/ECAR/MorePeopleNotJustMoreStuffDeve/163667 .

Garritano, J. R. and J. R. Carlson (2009). A subject librarian's guide to collaborating on e-Science projects. ISTL Issues in Science and Technology Librarianship, Spring 2009. Retrieved December 15, 2009, from http://www.istl.org/09-spring/refereed2.html

Gold, A. (2007a). Cyberinfrastructure, data, and libraries, Part 1: A cyberinfrastructure primer for librarians. D-Lib Magazine, 13 (9/10), September/October 2007. Retrieved February 6, 2010, from http://www.dlib.org/dlib/september07/gold/09gold-pt1.html

Gold, A. (2007b). Cyberinfrastructure, data, and libraries, Part 2: Libraries and the data challenge: Roles and actions for libraries, D-Lib Magazine, 13 (9/10), September/October 2007. Retrieved February 6, 2010, from http://www.dlib.org/dlib/september07/gold/09gold-pt2.html

Gray, J. and A. Szalay (2004). Where the rubber meets the sky: Bridging the gap between databases and science. Technical Report, Microsoft Research, Microsoft Corporation (MSF-TR-2004-110). Published in IEEE Data Engineering Bulletin, 27(4) 3-11, Dec. 2004. Retrieved April 4, 2010, from http://research.microsoft.com/apps/pubs/default.aspx?id=64540

*Gray, J. et al. (2009). The fourth paradigm: Data-intensive scientific discovery. Microsoft Research. Part 4: Scholarly Communication (essays by Clifford Lynch, Paul Ginsparg, Herbert Van de Sompel and Carl Lagoze, John Wilbanks, and others. Retrieved November 17, 2009, from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Green, T. (2009), We need publishing standards for datasets and data tables, OECD Publishing White Paper, OECD Publishing. Retrieved January 30, 2010 from
http://dx.doi.org/10.1787/603233448430

Greenberg, J. (2009). Metadata research supporting the Dryad data repository. Presentation given at Cornell University, April 17, 2009. Retrieved January 24, 2010 from
http://ecommons.library.cornell.edu/handle/1813/12247.

Griffin, S. M. (1998). NSF/DARPA/NASA Digital Libraries initiative: A program manager's perspective. D-Lib Magazine, July/August 1998. Retrieved January 30, 2010 from
http://www.dlib.org/dlib/july98/07griffin.html

*Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. Library Trends, 57(2), 280-299. Retrieved January 30, 2010, from
http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html

Hunt, K. (2004). The challenges of integrating data literacy into the curriculum in an undergraduate institution. IASSIST Quarterly 12, Summer/Fall 2004. Retrieved December 15, 2009, from http://iassistdata.org/publications/iq/iq28/iqvol282_3hunt.pdf

*International Journal of Data Curation. Retrieved December 10, 2009 from
http://www.ijdc.net/index.php/ijdc

Jones, E (2009). Reinventing science librarianship: Themes from the ARL-CNI forum. Research Library Issues, 262 (Feb. 2009). Retrieved January 24, 2010 from
http://www.arl.org/events/fallforum/forum08/index.shtml]

Key Perspectives (2010). Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study. Digital Curation Centre. Retrieved January 30, 2010, from http://www.dcc.ac.uk/scarp.

Lagoze, C. (2003). NSF DL Position paper. Prepared for NSF post digital library futures workshop, June 15-17 2003. Retrieved January 30, 2010 from http://www.sis.pitt.edu/~dlwkshop/paper_lagoze.html

Lankes, D. (2009). Cyberinfrastructure facilitators: New approaches to information professionals for E-Research" Oxford e-Research'08 Conference, Oxford, UK. Retrieved November 13, 2009, from http://quartz.syr.edu/rdlankes/blog/?p=542.

Lee, J. W., et al. (2009). DataNet: An emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning. AIChE Journal 55 (11), p. 2757-2764 Retrieved January 23, 2010 from http://dx.doi.org/10.1002/aic.12085

Lewis, D. W. (2007). A strategy for academic libraries in the first quarter of the 21st century, College & Research Libraries, 68(5):418-434, September 2007. Retrieved December 10, 2009 from http://idea.iupui.edu/dspace/handle/1805/1592

Lougee, W. and N. Rambo (2008). Library engagement with E-Science. ARL, Coral Gables, May 2008. Retrieved January 24, 2010 from http://www.arl.org/bm~doc/mm152_lougee.pps

Lyon, L. (2003). eBank UK: Building the links between research data, scholarly communication and learning. Ariadne 36, July 2003. Retrieved January 30, 2010 from http://www.ariadne.ac.uk/issue36/lyon/intro.html

Madnick, S., M. Smith, and K. Clopeck (2009). How much information? Case studies on scientific research at MIT (HMI case studies). June 2009. Global Information Industry Center, University of California, San Diego. Retrieved January 23, 2010 from http://hmi.ucsd.edu/pdf/MIT_case_studies_combination.pdf [See also 1 hr. webinar, at http://hmi.ucsd.edu/pdf/webinar/lib/playback.html, recorded July 2009, retrieved January 23, 2010]

McCue, J., and J. Corson-Rikert (2007). Engaging and connecting faculty: Research discovery, access, re-use, and archiving. Project Briefing, CNI, Spring 2007. Retrieved December 10, 2009, from http://www.cni.org/tfms/2007a.spring/abstracts/PB-engaging-mccue.html

MIT Libraries (2008). Data management and publishing (research guide). Retrieved January 30, 2010, from http://libraries.mit.edu/guides/subjects/data-management/index.html

MIT Libraries. (n.d.). Scholarly Publication – MIT Libraries » Open Data. Retrieved November 11, 2009, from http://info-libraries.mit.edu/scholarly/open-access-initiatives/faq/open-data/

Mullins, J. (2007). Enabling international access to scientific data sets: Creation of the Distributed Data Curation Center (D2C2). Retrieved November 18, 2009, from http://docs.lib.purdue.edu/lib_research/85/

National Academies of Sciences (2007). Designing cyberinfrastructure for collaboration and innovation. January 29-30, 2007. Retrieved December 10, 2009, from http://cyberinfrastructure.us/resources.htm
Selected papers also published as issue of First Monday. Retrieved December 10, 2009, from http://www.firstmonday.org/issues/issue12_6/

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. Retrieved December 11, 2009, from http://www.nap.edu/catalog.php?record_id=12615)

National Science Board (2005). Long-lived digital data collections: Enabling research and education in the 21st century. Retrieved January 24, 2010, from http://www.nsf.gov/pubs/2005/nsb0540/

National Science Foundation and Joint Information Systems Council (2003). Wave of the future: NSF post digital library futures workshop. June 15-17, 2003, Chatham, Massachusetts. Retrieved January 30, 2010, from http://www.sis.pitt.edu/~dlwkshop/

National Science Foundation (2007a). Sustainable digital data preservation and access network partners (07-601, DataNet). Retrieved December 30, 2009, from http://www.nsf.gov/publications/pub_summ.jsp?WT.z_pims_id=503141&ods_key=nsf07601

National Science Foundation (2007b). Informational meeting for potential applicants for DataNet funding, November 8, 2007. Retrieved January 24, 2010 from http://www.nsf.gov/news/news_summ.jsp?cntn_id=110392

Pennington, D. (2008). Cross-disciplinary collaboration and learning. Ecology and Society 11 (2). Retrieved January 30, 2010 from http://www.ecologyandsociety.org/vol13/iss2/art8/

*Pepe, A., Mayernik, M., Borgman, C., and Van de Sompel, H. (2009) From artifacts to aggregations: Modeling scientific life cycles on the Semantic Web. Journal of the American Society for Information Science and Technology. Retrieved February 6, 2010, from http://dx.doi.org/10.1002/asi.21263.

Renear, A., and Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. Science, 325(5942), 828-832. Retrieved January 25, 2010, from http://www.sciencemag.org/cgi/content/abstract/325/5942/828

*Research Data Management discussion list (RESEARCH-DATAMAN). "List to discuss data management issues arising in and from research projects in Higher Education communities, in the UK and internationally, established by the Digital Curation Centre on behalf of the JISC." https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=RESEARCH-DATAMAN (retrieved April 4, 2010).

Rusbridge, C. (2005). Information life cycle and curation (Presentation). Retrieved January 30, 2010 from http://www.dcc.ac.uk/docs/dcc-life-cycle.ppt

Rusbridge, C., et al. (2005). The Digital Curation Centre: A vision for digital curation. In: proceedings of Local to Global Data Interoperability – Challenges and Technologies, 20-24 June 2005, Sardinia, Italy. Retrieved December 3, 2009, from http://eprints.erpanet.org/82/

*Rusbridge, C. (2007 - ) Digital Curation Blog. Retrieved December 3, 2009, from http://digitalcuration.blogspot.com/

Sallans, A., S. Lake, K. Miles, and R. Pappert. Data services in the library? Challenges and successes in the present and future. Poster presented at 2009 ACRL conference, STS Section meeting, Big Science, Little Science, E-Science. Retrieved January 24 from http://www.ala.org/ala/mgrps/divs/acrl/about/sections/sts/conferences/poster2009sallans.pdf

Schmidt, M., and R. Reznik-Zellen (2009). E-Science @ U Mass. Poster presented at 2009 ACRL conference, STS Section meeting, Big Science, Little Science, E-Science. Retrieved January 24 from http://www.ala.org/ala/mgrps/divs/acrl/about/sections/sts/conferences/poster2009schmidt.pdf

Shearer, E., (2009). Research data: Unseen opportunities. An awareness toolkit commissioned by the Canadian Association of Research Libraries. Retrieved January 24, 2010, from http://www.carl-abrc.ca/about/working_groups/pdf/data_mgt_toolkit.pdf

Shreeves, S., and M. Cragin. (2008). Introduction: Institutional repositories: Current state and future. Library Trends 57 (2), Fall 2008. Retrieved December 15, 2009, from http://www.ideals.illinois.edu/handle/2142/10679

Smith, M. (2009). Webinar: Bringing research data into the library: Expanding the horizons of institutional repositories. Tuesday, November 10, 2009. Retrieved November 11, 2009, from http://www.ala.org/ala/mgrps/divs/alcts/confevents/upcoming/webinar/ir_series.cfm#6

Spengler, S. (2009). DataNet: A Sustainable Digital Data Preservation Network. Presentation to CENDI, March 10, 2009. Retrieved February 7, 2010, from http://www.cendi.gov/minutes/pa_0309.html#spengler or http://www.cendi.gov/presentations/CENDI_03-10-09_Spengler_NSF_DataNet.pdf

*Steinhart, G., Saylor, J., and Westbrooks, E. L. (2008). Digital research data curation: Overview of issues, current activities, and opportunities for the Cornell University Library. Retrieved November 16, 2009, from http://ecommons.cornell.edu/handle/1813/10903.

Steinhart, G., D. Dietrich, and A. Green (2009). Establishing trust in a chain of preservation: The TRAC checklist applied to a data staging repository (DataStaR). D-Lib Magazine 15 (9/10), September/October 2009. Retrieved January 24, 2010 from http://www.dlib.org/dlib/september09/steinhart/09steinhart.html

Swan, A., and S. Brown (2008). Skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. Retrieved November 22, 2009, from http://eprints.ecs.soton.ac.uk/16675

*Van Horik, Rene (2008). "Data curation," Chapter 5 in: A DRIVER'S guide to European repositories. Kasja Weenink and Leo Waaijers, Amsterdam University Press. Retrieved as PDF April 4, 2010 from http://dare.uva.nl/document/93898

Van de Sompel, H. and C. Lagoze (2009). All aboard: Toward a machine-friendly scholarly communication system. In: The Fourth paradigm: Data-intensive scientific discovery, Microsoft Research. Retrieved February 6, 2010, from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part4_sompel_lagoze.pdf

Whitlock, M. et al. (2010). Data archiving (editorial). The American Naturalist 175, pp. 145–146. Retrieved January 23, 2010 from: http://www.journals.uchicago.edu/doi/abs/10.1086/650340

Witt, M. and J. R. Carlson (2007). Conducting a data interview. Poster presented December 2007 at Digital Curation Conference. Retrieved December 20, 2009, from http://www.dcc.ac.uk/events/dcc-2007/programme/all-posters-demos

Witt, M. (2008). Institutional repositories and research data curation in a distributed environment. Library Trends 57 (2), Fall 2008, p. 191-201. Retrieved December 10, 2009, from http://muse.jhu.edu/journals/library_trends/v057/57.2.witt.html
Also available at: http://docs.lib.purdue.edu/lib_research/104/

Witt, M. (2009). Eliciting faculty requirements for research data repositories. Fourth International Open Repositories Conference (OR2009), Atlanta, GA, May 18-21, 2009. Retrieved January 5, 2010, from http://smartech.gatech.edu/dspace/handle/1853/28509

Witt, M., Carlson, J., Brandt, D.S., & Cragin, M.H. (2009). Constructing data curation profiles. International Journal of Digital Curation, 4(3), December, 2009. Retreived February 6, 2010 from http://www.ijdc.net/index.php/ijdc/article/view/137/165

Yakel, E., Conway, P., and Krause, M. (2009). Thinking like a digital curator: Creating internships in the cognitive apprenticeship model. DigCCurr 2009 proceedings, 7-11. Retrieved February 6, 2010, from http://www.slideshare.net/mghetu/digccurr-presentation-final?type=presentation

**Acknowledgements:**

This paper was originally developed for a presentation in December 2009 as part of a panel at the annual meeting of the American Geophysical Union. I thank Karen Baker and Cyndy Chandler for their generous invitation to participate in the panel, and NSF for their support (NSF/HSD #0433369). I also wish to thank several colleagues who provided thoughtful and informative additions, corrections, and editing suggestions to an earlier version of this paper, as well as encouragement: Karen Baker, Melissa Cragin, Tracy Gabridge, Marisa Ramirez, and Jeanine M. Scaramozzino. Corrections and comments are welcome at: anna.calif@gmail.com or agold01@calpoly.edu