2010

Review of Research Data Curation Practices and Attitudes of Stakeholders

Suzie Allard
University of Tennessee - Knoxville, sallard@utk.edu

Elizabeth Allen

Christina Murray

Robert J. Sandusky
University of Illinois at Chicago, sandusky@uic.edu

Review of Research Data Curation Practices and Attitudes of Stakeholders

Suzie Allard, Elizabeth Allen, Christina Murray and Bob Sandusky
2010

Advances in technology have allowed researchers to create large amounts of reusable digital data sets. This phenomenon, e-Research, encompasses not only innovative forms of scholarship in the sciences, but the humanities and social sciences as well (Lynch 2008). While digital data benefits research by permitting new types of problems to be addressed, increases ability to collaborate across disciplines and institutions, and allows for replication of previous results, it is easily lost for future use unless action is taken to manage it from its inception.

Data curation is the process of managing digital data from the moment of its creation so that it can be accessed, understood and potentially re-used in the future (Lord and Macdonald 2003). The steps necessary to curate research data have been outlined through the Data Curation Centre's Curation Lifecycle Model (Digital Curation Centre n.d.). Activities for data management include planning for the creation of data, describing the data using standardized metadata, housing the data set in a repository, creating data management plans, migrating objects as necessary to overcome media decay and format obsolesence, and appraising for selection and deselection. Long-term management of digital data sets is still in its infancy, and numerous issues such as what kind of expertise is needed to properly curate data and how they should be financially provided for into the future when grant money ceases are not fully resolved (Lynch 2008, Lyon 2007).

Due to the importance of digital data for research, many librarians have discussed what their future role of librarians may be at academic institutions (Council on Library and Information Resources 2008, ARL 2006, Hey and Hey 2006). The increase in digital research data has created an opportunity for library involvement in data curation (Gold 2007). In response, libraries, universities, and other organizations, have surveyed either researcher's data curation needs and current practices or library involvement and policies in data curation. However, little research has been conducted on individual librarians to assess current activities in data curation, and, in particular, their attitudes towards it.

Surveys of Library User's Needs

Though librarian attitudes toward research data curation have rarely been assessed, libraries and other organizations have frequently conducted assessments of library users' perceptions and needs in this regard. A few such studies are described below, a list which is not intended to be exhaustive. ARL's 2009 E-Science Survey website lists three surveys of data needs conducted by ARL member libraries: at University of Oregon, University of Wisconsin (which conducted two separate assessments), and a joint project by Purdue and the University of Illinois at Urbana-Champaign. Also, in 2006, the Provost at Yale University administered a similar survey to their faculty. Each addressed research data to varying degrees.

As a part of its Science Data Audit, the University of Oregon Libraries continuously surveys UO researchers to assess the datasets they generate, in order to "identify one or more potential partners from the sciences for a pilot data curation project, and/or collaboration on grant proposals with data curation components or infrastructure" (Westra, 2010). In 2008, the University of Wisconsin Madison released the Research Data Management Study Group's (RDMSG) Summary Report, the result of eight interviews with twenty-one researchers across campus. The RDMSG grew out of the Scholarly Assets Management Initial Exploratory Group (SAMIEG), which conducted focus group discussions and interviews aimed at determining the nature of data currently being produced, how it is being used and funded, and what researchers would like to do with it in the future. SAMIEG's interviewees included a map and GIS librarian and a curator at Wisconsin's Herbarium Library (Simpson et al., 2007, p. 23). The Distributed Data Curation Center in the Libraries at Purdue University and the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign are collaborating on the Data Curation Profiles Project, which began in 2007 and has produced case studies of researchers' data practices, including scientific workflow and data sharing.

Some assessments on the research community's data curation needs are undertaken by organizations other than libraries. Yale University's Provost created an online survey to ascertain faculty member's feelings of importance on various issues of cyberinfrastructure to their work and what should be a priority for Yale University as a whole. For both the faculty member's work and priority for Yale University statements, faculty rated "easier electronic access to scholarly materials" highest. Interestingly, the statement "ensuring the preservation of my scholarly digital output (datasets, research notes, e-prints)" was ranked as 11th out of 19 for importance to individual faculty and 15th as a priority for the University.

Ultimately, these institutional assessments serve the practical purpose of helping libraries and universities plan to meet user needs. Also, they give some insight into the questions preoccupying library working groups. It is also worth noting who conducted these assessments. Some were the work of librarians specializing in research data, such as UO's Science Data Services Librarian. Others involved librarians in a range of professional roles. For example, the studies at Wisconsin involved digital services and digital repository librarians, and interviewers for the Digital Curation Profiles Project included subject librarians in fields such as health science, agricultural science, chemistry, geology, geographic information systems, anthropology and sociology. These surveys help librarians assess their user's needs, but with the exception of Wisconsin, not their attitudes towards data curation as an activity for librarian involvement.

Surveys of Libraries

Libraries were the object of study in ARL's 2009 E-Science Survey (ARL 2009a). ARL's investigations into cyberinfrastructure and libraries began in 2004 with its "E-Research and Supporting Cyberinfrastructure: A Forum to Consider the Implications for Research Libraries & Research Institutions," held only a year after the publication of the Atkins report. ARL's e-science task force began in 2006, which released "To Stand the

Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering" the same year. ARL's efforts focused on issues of education, workforce development, collaborations with other organizations, and policy development. The task force became a working group, which then carried out the survey to discover how member libraries were participating in their institutions' e-science initiatives and identify some of the pressure points that affected these initiatives and library participation in them.

The survey was primarily descriptive in nature, with an ancillary goal of uncovering models of e-science services. Its scope extended beyond the library, to e-science planning efforts and data support services at the institution as a whole, whether institution-wide, confined to individual units, or a hybrid of the two. Subsequent questions focused on library participation: who at the library is responsible for data support, whether an individual, a committee, or a department, along with the position titles, years of experience, and place in the library hierarchy of those responsible. Respondents detailed services the library offers, enumerated as "reference/consultation activities" ("finding and using available technology infrastructure and tools, finding relevant data, developing data management plans, developing tools to assist researchers"), informative web sites, workshops for researchers, advice on policy issues, and infrastructure ("data storage, tools for data analysis, virtual community support"). How the library has developed its workforce for e-science participation, and how it has collaborated with other units within and beyond the institution, round out the survey's areas of inquiry.

ARL's survey has a clear emphasis on institutional structures rather than on the individuals that comprise those structures. The survey did ask for details of up to three positions at the library that engage with research data, including terms of contract, title, degrees held, and job descriptions. Though it is a survey of libraries rather than librarians, it was by necessity completed by a librarian. One of its final questions asked for "pressure points for your institution and your library related to e-science support or e-research more broadly" (p. 25), and as pressure points are a matter of opinion, answers to this question cannot help but reflect the attitudes of the individual respondents. Aside from these points, however, attitudes are not touched upon.

At the ARL membership meeting on October 14, 2009, Wendy Lougee presented the initial findings (ARL 2009b). Fifty-two respondents reported on the institutional structures that dealt with research data, 75% of whom had some form of e-science support at their institutions, either now or planned to be provided in the future. Eighty-six percent of libraries collaborated with other institutional units to participate in data curation. A final report is expected to be released later in 2010.

Although not a survey on library involvement in data curation, Cornell University Library (CUL) created a working group in 2006 to conduct an environmental scan of research data curation activities at Cornell University and beyond, including other academic libraries, and to make recommendations on possible opportunities for CUL involvement in research data curation. The working group, Data Working Group (DaWG), published the outcomes of their environmental scan and recommendation in a white paper (Steinhart et al. 2008). DaWG determined that a few U.S. academic libraries

were engaged in data curation of research data. Staff at the Sheridan Libraries of John Hopkins University curate astronomical data (Choudhury et al. 2007) and are involved in many other projects through the Digital Research and Curation Center (http://ldp.library.jhu.edu/dkc). Also, Purdue University Libraries hired a director, data research scientist, and interdisciplinary research librarian to manage the new Distributed Data Curation Center (http://d2c2.lib.purdue.edu/) which operates a repository, e-Data, specifically for research data (http://www4.lib.purdue.edu/lcris/edata/#4). The University of Washington Libraries created and hired for the position of Director of Cyberinfrastructure Initiatives and Special Assistant to the Dean of University Libraries for Biosciences and e-Science. At the time of publication, CUL was already involved in some research data curation activities. Their institutional repository, E-common, holds some small data sets, Cornell University Geospatial Information Repository holds geospatial data and metadata for New York State, and the Library along with five USDA agencies provide access to electronic data on U.S. and international agriculture. In addition, CUL was expecting to participate in additional research data curation projects including DataStaR, an NSF-funded project to create "an institutionally-based data staging repository whose function is to facilitate the documentation and transmission of research data sets from a variety of disciplines to domain-specific repositories and/or institutional repositories" (pg. 16). From DaWG's research, it is clear that some libraries, including Cornell University Library, are beginning to participate in research data curation, with an increase in new positions and centers to meet this need. As noted in the white paper, all these projects involve collaboration between the library and other organizations, both within and beyond its parent university.

The DaWG's environmental scan led them to conclude that research data curation faces six categories of issues: financial stability, appraisal and selection, digital preservation, intellectual property, confidentiality, and participation by data owners. These issues, in turn, guided their recommendations for activities CUL should be embarking on to further their involvement in research data curation. CUL should actively seek out partnerships with other organizations, provide data management and archiving services to support Cornell researchers, determine Cornell University needs and build local infrastructure and policy as needed, identify new skills required for CUL staff and promote staff development, and form a dedicated executive data curation group. By conducting a thorough literature review, combined with an environmental scan of research data curation in a variety of organizations, this study provides a benchmark of current research data curation within Cornell University Library and other libraries.

These two surveys of library activity focus on what initiatives have been taken or are being developed; what could be developed in the future is only touched upon briefly, either as pressure points or training needs. Also, these surveys do not investigate the impetus behind library data activities or the conditions that encouraged them, which might allow libraries considering such programs to estimate their chances of success. Finally, with a focus on library structures these studies do little to examine the current readiness or training needs of the individual librarians who enact these structures. The next section will discuss some surveys of librarians and research data.

Surveys of Librarians

Studies of librarians and research data have most frequently been conducted in Europe. A UK study purported to study the perceptions of librarians regarding research data, but in fact focused on librarian activities. In 2005, the Consortium of University Research Libraries (CURL, now RLUK) partnered with the Society of College, National and University Libraries (SCONUL) to form the CURL/SCONUL Joint Task Force on e-Research, "to investigate the impact of e-Research on the academic library sector, including developing training to assist library and information professionals extend their knowledge and skills" (Martinez, 2007, p. 3). One product of this task force was an online survey that targeted librarians in liaison, researcher support, and data preservation roles, and was published in 2007 as "e-Research Needs Analysis Results." Although one of the objectives of the survey was "to better understand the perceptions and expectations of library staff in relation to Library involvement" in e-research (p. 3), examination of the survey instrument reveals that data on few perceptions were elicited. The bulk of the survey concentrated on whether various research data services (grid technologies, metadata development) were "established," "current area of development," or "under discussion" at the respondent's institution, with a fourth option for "not aware." The language of the report concludes, from the high number of "not aware" responses for "support for data web developments" and "development of metadata associated with primary research data," that there was "a significant high unawareness of any discussion" of these topics (p. 9). However, from the survey alone it is impossible to determine whether a "not aware" response indicates unawareness or the actual absence of such a service at the respondent's institution. The only question that surveyed perceptions asked respondents to prioritize subjects they would like to see included in training workshops for librarians. The top two responses were "research skills development in relation to e-research and how to address this" and "researchers' views of Library involvement in e-research" (p. 14). The latter suggests that some issues regarding library-researcher relations are worth exploring further. Indeed, in the free text comments to this question, one respondent wrote "How to persuade researchers that librarians have skills they can make use of!" (p. 15).

Another UK survey of librarians co-sponsored by CURL and the Research Information Network, "Researchers' Use of Academic Libraries and Their Services," addressed a much wider range of issues than the CURL-SCONUL study, and only touched on the topic of research data. When asked to identify the core roles for librarians in five years time, 32% of the 300 surveyed librarians judged "manager of datasets from e-science/grid projects" to be a future core role, with another 35% designating it an ancillary role. Notably, of all the roles listed it aroused the most uncertainty, with 18% selected "don't know."

In 2008, Alma Swan and Sheridan Brown presented their report "The Skills, Role, and Career Structure of Data Scientists and Curators: An Assessment of Current Practices and Future Needs" to the Joint Information Systems Committee (JISC). Information gathered through fifty-seven semi-structured interviews with data scientists, librarians, and educators allowed Swan and Brown to make recommendations concerning the development of data scientists and curators to best aid researchers (Swan and Brown, 2008, p. 1). Though librarians are not explicitly named in the title of the report, "data

librarians," those "originating in the library community, trained and specialising in the curation, preservation, and archiving of data" (1), fill one of the four data-related roles the authors list. Much of the study focuses on data scientists, another of the four data-related roles, to the exclusion of librarians, though the authors do propose three key roles for data librarians: "increasing data-awareness amongst researchers; providing archiving and preservatin [sic] services for data within the institution and through institutional repositories; and developing a new professional strand of practice in the form of data librarianship" (2). Interviewed librarians reported that they are "increasingly being approached by researchers for advice [sic] practical help with data management" (25); the methodology of the study does not allow for estimation of this increasing rate. The report lists several reasons, presumably culled from interviews, as to why there is a shortage of skilled data librarians: lack of information among LIS students about the field and its job prospects, a paucity of suitable internships, and the need for subject expertise in science (27). How much domain knowledge is required to work with research data remains to be decided, but as a perceived requirement it may form a barrier to librarians who might engage in data curation.

PARSE.Insight, funded by the European Union, conducted a survey of researchers, data managers, publishers, and researcher funders, with the objective of defining "the needs for an e-Science infrastructure for long-term availability of research data" (Kuipers & Van der Hoeven, 2009, p. 4). "Data managers" were defined as "professional with a clear responsibility for the preservation of research data and publications," found at "research libraries, data centres, archives, and other data management organizations" (p. 37). Although the respondents were not exclusively librarians, 73% were employed at libraries, most likely because the mailing list of the Association of European Research Libraries (LIBER) was the distribution channel with the most responses, at 45% of the total (p. 38). Of the 273 respondents, more than three-quarters were European.

This survey investigated both perceptions of data preservation as well as the current state of preservation at respondents' institutions. Data managers were asked to rate, on a four-point scale ranging from "very important" to "not important," the reasons to preserve research data. The three reasons rated with the most "very important" and "important" responses include: if research is publicly funded, the results should become public property and therefore be properly preserved; preservation of research data will stimulate the advancement of science (new research can build on existing knowledge); and preservation allows for re-analysis of existing data (pg. 39). Similarly, respondents rated the following three threats to preservation with the most "very important" and "important" responses: lack of sustainable hardware, software or support of computer environment may make the information inaccessible; users may be unable to understand or use the data, e.g. the semantics, format or algorithms involved; and the current custodian of the data, whether an organization or project, may cease to exist at some point in the future (pg 40). Respondents were asked whether international infrastructure for data preservation and access should be built to help guard against these threats, to which sixty percent answered yes.

All four studies on librarians provide insight into current and future library involvement in data curation and both researcher, and to a much smaller extent, librarian attitudes towards data curation as a role for libraries. The study "Researchers' Use of Academic Libraries and Their Services" illuminates the uncertainty librarians have in their role as data set managers, while "The Skills, Role, and Career Structure of Data Scientists and Curators" study is a third-party recommendation for librarians to be involved in data curation with archiving and preservation services that they provide to their community, particularly "small science" data sets. Other curation roles should be performed by data scientists and managers. This study also highlights some of the barriers to involvement including a lack of properly trained librarians. The PARSE.Insight project differed from the other two studies because it surveyed librarians who were already engaged in the management of research data sets and asked attitudinal questions, such as what are important motivations for preserving data sets and what are the obstacles to doing so. Given the thin number of surveys of librarians, particularly of those in the United States, and the large potential for librarian involvement in data curation, more surveys should be conducted to assess current librarian participation in data curation and their attitudes towards it.

Conclusion

Who will be facilitating data curation on research data and in what capacity is not a foregone conclusion. A review of surveys conducted on researcher's data needs, current library and librarian practices demonstrate that some libraries are already involved, but there is scant information from individual librarians on their attitudes towards data curation. While the ARL's E-Science Survey assessed data curation participation at the library-level, it is also important to understand librarian attitudes, including what motivates and hinders them to contribute to research data sets, and their level of individual participation. This information is necessary to provide 1) baseline data on attitudes and 2) identify opportunities/barriers for future librarian participation in the curation of research data sets.

References

Association of Research Libraries (2006). To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering. Washington, DC: Author.

Association of Research Libraries : E-Science Survey Resource Page. (2009a). Retrieved June 14, 2010, from http://www.arl.org/rtl/eresearch/escien/esciensurvey/surveyresearch.shtml#instdatres.

Association of Research Libraries (2009b). *ARL Survey on E-Science and Data Support: Initial Findings.* Retrieved June 28, 2010 , from http://www.arl.org/resources/pubs/mmproceedings/155mm-proceedings.shtml#esci

Choudhury, S., DiLauro, T., Szalay, A., Vishniac, E., Hanisch, R., Steffen, J., ...Plante, R. (2007). Digital Data Preservation for Scholarly Publications in Astronomy. *International Journal of Digital Curation* 2(2), 20-30.

Council on Library and Information Resources (2008). No Brief Candle: Reconceiving Research Libraries for the 21st Century. Washington DC: Author.

Digital Curation Centre. (n.d.). *DCC Curation Lifecycle Model*. Retrieved June 20, 2010 from http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Gold, A. (2007). Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *D-Lib Magazine* 13, (9/10). Retrieved June 21, 2010 from http://www.dlib.org/dlib/september07/gold/09gold-pt2.html.

Hey , T. & Hey, J. (2006). "e-Science and its implications for the library community", *Library Hi Tech*, Vol. 24 Iss: 4, pp.515 – 528.

Kuipers, T. & Van der Hoeven, J. (2009). Survey Report (D3.4). Didcot, UK: PARSE.Insight. Retrieved June 23, 2010 from http://www.parse-insight.eu/publications.php#d3-4.

Lord, Philip & Macdonald, Alison. (2003). *E-science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*. Retrieved June 28, 2010 from http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf.

Lynch, C. (2008). The Institutional Challenges of Cyberinfrastructure and E-Research. *EDUCAUSE Review*, vol. 43, no. 6. Retrieved June 15, 2010 from http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/TheInstitutionalChallengesofCy/163264.

Lyon, L. (2007). Dealing with Data: Roles, Rights, Responsibilities and Relationships. Bristol: JISC. Retrieved June 14, 2010 from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report.pdf.

Martinez, L. (2007). The e-Research needs analysis survey report. London: CURL/SCONUL Joint Task Force on e-Research. Retrieved June 23, 2010 from http://www.rluk.ac.uk/files/E-ResearchNeedsAnalysisRevised.pdf.

Purdue University Library. 2010. "Data Curation Profiles." Retrieved June 14, 2010, from http://wiki.lib.purdue.edu/display/dcp/Data+Curation+Profiles

RIN and CURL. (2007). Researchers' Use of Academic Libraries and their Services: A report commissioned by the Research Information Network and the Consortium of Research Libraries. London: authors. Retrieved June 23, 2010 from http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/researchers-use-academic-libraries-and-their-serv

Simpson, M., Cheetham, J., Gorman, P., Herr-Hoyman, D., Larson, E., Salo, D., et al. (2007). Summary Report of the Scholarly Assets Management Initial Exploratory Group. Retrieved May 26, 2010, from http://digital.library.wisc.edu/1793/21443.

Steinhart, G. et al. (2008). Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library. Retrieved June 14, 2010 from http://hdl.handle.net/1813/10903.

Swan, A. & Brown, S. (2008). The Skills, Role And Career Structure Of Data Scientists And Curators: An Assessment Of Current Practice And Future Needs. Truro, UK: Key Perspectives Ltd. Retrieved June 23, 2010 from http://eprints.ecs.soton.ac.uk/16675/.

Westra, B. (2010). Science Data Services - Data Inventory. *UO Libraries*. Retrieved June 14, 2010, from http://libweb.uoregon.edu/faculty/SciDataAudit.html.