8th International Digital Curation Conference
January 2013

The Problematic Future of Research Data Management: Challenges, Opportunities, and Emerging Patterns Identified by the DataRes Project

Martin Halbert

Dean of Libraries,

University of North Texas

Abstract

This paper describes findings and projections from a project that has examined emerging policies and practices in the United States regarding the long-term institutional management of research data. The DataRes Project of the University of North Texas (UNT) studied institutional transitions taking place during 2011-2012 in response to new mandates from U.S. governmental funding agencies requirements for research data management plans to be submitted accompanying grant proposals. Additional synergistic findings from another UNT project termed iCAMP will also be reported briefly.

This paper will build on these data analysis activities to discuss conclusions and prospects for likely developments within coming years based on the trends surfaced in this work. Several of these conclusions and prospects are surprising, representing both opportunities and troubling challenges for not only the library profession but the academic research community as a whole.

The 8th International Digital Curation Conference will take place on 14–16 January 2013 in Amsterdam. Please ensure you use the guidance in this template to produce your paper. Please submit your paper in one of the following formats: Microsoft Word (.doc), Open Office (.odt) or Rich Text (.rtf).
http://www.dcc.ac.uk/events/idcc13/submissions

Introduction

Most governmental funding agencies in the United States have now mandated or are in the process of mandating data management plans as requirements for submitting research grant applications. Prominent U.S. agencies that have led this trend include the National Science Foundation (NSF), the National Institutes of Health (NIH), and the National Endowment for the Humanities (NEH). As a result, research universities across the country are now struggling to develop consistent policies and programmatic implementations for institutional data management functions.

Research libraries and library and information science (LIS) programs in particular are scrambling to respond to these new requirements and to understand emerging requirements for curricula and training for both students and working information professionals. Recent white papers (ARL, 2006, 2007) and task force reports (Berman et al., 2010) provide evidence that there is an acute need for research that will inform this process of curriculum and training development; research that documents the emerging patterns in data management policies, and the expectations of major stakeholders in the research cycle regarding data management roles, responsibilities, and professional training and preparation for those taking on data management responsibilities.

The DataRes Project (http://datamanagement.unt.edu/datares) is investigating the emerging institutional policy responses of U.S. universities in response to these new grant application requirements for research data management plans. The project is termed DataRes as a shorthand mnemonic for the broad themes concerning research data which it is examining. DataRes is based at the University of North Texas (UNT), and was funded by a 2011 grant of US$ 226,786 from the IMLS 21st Century Librarian (21CL) program. In the course of analysis, the DataRes project is also studying how the library and information science (LIS) profession can best respond to emerging needs of research data management in universities, and was paired with another IMLS 21CL 2011 grant of US$ 624,663 to UNT for the iCAMP project to assess educational needs and develop new shared curricula to train new LIS professionals seeking to fill data management positions.

The two projects, DataRes and iCAMP, represent a close collaboration between the UNT Libraries and the UNT College of Information. Both the DataRes and iCAMP projects have now completed their primary data analysis activities, and these findings will be published in an edited volume on data management trends by the Council on Library and Information Resources (CLIR) in 2013. While the emphasis of this paper is on the DataRes project, the data analysis and findings of both projects will be used to draw conclusions and projections of likely developments within the next 5-10 years based on the trends surfaced in this research.

Several of these conclusions and projections are surprising, and represent both opportunities and troubling challenges for not only the library profession but the academic research community as a whole. Before reviewing these conclusions and projections, a brief review will be provided of the two projects, their main goals, and their methodologies.

DataRes and Affiliated Projects

The genesis of the DataRes and iCAMP projects was driven by the overwhelming new emphasis on research data management that became acute during the latter half of the first decade of the 21st Century. The UNT Libraries and UNT College of Information have been close collaborators for years, and sought a way of mutually engaging in a constructive process of exploring the emerging landscape of data management needs. The result of the ensuing planning process was that two affiliated projects were envisioned: one project to document and analyse the emerging landscape of university responses to new data management needs (with an aim to clarify the role of libraries and information centers in this new landscape), and one project to develop new educational curricula for the next generation of information professionals serving to meet these needs. While this paper focuses primarily on the DataRes project, the synergistic findings and results of the iCAMP project will also be discussed along the way.

DataRes Project Background and Literature Review

The project was informed by a thorough literature review of the growing body of important work in the area of digital curation, and data management specifically. While there are already too many major works to even list here, there were several very significant studies of which the project particularly took note.

One was the JISC report by Alma Swan and company which investigated the skills and emerging roles of data scientists and curators (Swan & Brown, 2008). Swan's study helpfully contextualized the ambiguity of these emerging roles, educational preparation, and the complexity of institutional responses. Her prescriptive conclusion that "The role of the library in data-intensive research is important and a strategic repositioning of the library with respect to research support is now appropriate" (p. 2) is a good summary of a fundamental conviction that is held by the LIS professionals of UNT (and many other institutions, obviously), and is what motivated this work at UNT.

The Association of Research Libraries (ARL) completed a survey of their members entitled "E-Science and Data Support Services" in August 2010. (Soehner, Steeves & Ward, 2010) While this survey was primarily aimed at E-Science support in research libraries, it also had significant portions focusing on data support and management functions in university libraries. This survey of research libraries highlighted a number of key points (pp. 8-9, emphases added):

> '…additional information about incentives and policies for the use of centralized data centers would be a useful component to understanding and creating successful centralized services.
>
> 'The top three areas identified by survey respondents as pressure points include a lack of resources, difficulty acquiring the appropriate staff and expertise to provide e-science and data management or curation services, and the lack of a unifying direction on campus.

> 'This area is very important, but is much larger than a single institution. We need a national framework for addressing the management, re-use and preservation of scientific data.'

ARL also released an online guide entitled *Guide for Research Libraries: The NSF Data Sharing Policy*. (Hswe & Holt, 2010) This guide "investigates the role of libraries in data management planning, offering guidance in helping researchers meet the NSF requirement." The guide was similar to other planning tools from educational institutions (example: Monash University, 2009), which highlight practices developed primarily in the context of specific institutional repositories.

There were a small number of inter-institutional studies of this topic which also highlighted the need for systematic research efforts in institutional data management issues. The U.S. report *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age* (National Academy of Sciences, 2009) which affirms a "Data Integrity Principle" highlighting the centrality of research data safeguards, and the responsibility of researchers and other stakeholders in ensuring the integrity of research data. The UKDA document *Managing and Sharing Data: A Best Practice Guide for Researchers* (Van den Eynden et al., 2009) was perhaps the most informative document examined in the literature review, and one that very much conveyed the importance of developing organizational policies for data management.

However, what emerged from this literature review was that there was need for a systematic survey to document the specific institutional responses to the new agency grant submission requirements organizational data management policies and practices that were rapidly emerging in response. The DataRes project was designed to do precisely this for the U.S. context.

DataRes Project Methodology and Findings

Several methods were used to first study and understand the data management policy requirements of the federal granting agencies, and then survey and assess the institutional responses to these requirements.

Analysis of Agency Data Management Plan Guidance

The project team first sought a better understanding of the requirements concerning data management plans issued by three U.S. federal grant-issuing agencies: the NSF, NIH, and NEH-ODH. The Institute of Museum and Library Services (IMLS) was not considered, because while the agency requires applicants to complete a questionnaire ("Specifications for Projects that Develop Digital Products"), it does not offer guidance for researchers beyond the content of the questionnaire.

As an initial heuristic approach to understanding the agency requirements, the project team extracted the various texts which constituted guidance or requirements for researchers concerning data management plans from the NSF, NIH, and NEH-ODH, and input them into two very different textual analysis systems. The first was the website *wordle.net*, a popular word cloud site which generates simple visualizations of texts. The project team limited the cloud visualizations to the top 100 words in each document as an initial way of understanding the relative

prominence of different terms in the respective agency statements. The word clouds were useful as a first concept mapping of the agency statements and suggested that further analysis based on text mining could be fruitful.

The same texts were then analysed using the Voyant suite of tools for lexical analysis developed by *hermeneuti.ca*, which provided additional capabilities for analysing word frequency, word associations, vocabulary density, distinctive word counts, peaks and trends in frequency for the individual documents. The project team then applied a Taporware stop words filter provided by Voyant to eliminate commonly used words like conjunctions and articles. These quantitative analyses served to target and enrich subsequent close readings of these agency statements.

The textual analysis showed that there was much more variation than originally anticipated between the different federal agency requirements for data management plans. The "Final NIH Statement on Sharing Research Data" is comprised of 869 words, with the most frequently used words in the document being "data" (29 uses) and "sharing" (26 uses). In every instance of "sharing", the word "data" appears either adjacent or within three words. This correlation is a strong indication of the culture of data sharing which the NIH requirement seeks to fosters. The places and manner in which the agency acronym "NIH" (16 uses) appears underscored the agency's authority as an arbiter of research data practice in the community it both serves and oversees.

The NSF's guidance to researchers (Award and Administration Guide. Chapter VI.D.4) is the smallest of the statements examined; at only 350 words it is, in fact, a small component of a larger document. Yet, it has the greatest vocabulary density (greatest instance of unique words). "NSF" appears 7 times in the document, "investigators" 5 times, and "grantees", "dissemination", and "results" each 4 times. There is no preponderance of usage of any of the key terms ("data", "management", or "sharing") as in the other agency guidance. "Data" in fact appears only 3 times. "NSF" occurs 3 times paired directly with "grants". Upon close reading, the focus (such as it is) appears to be on the initial assertion of authority of the granting agency on this emerging area of focus. Interestingly, each Directorate within the NSF provides supplemental guidance for applicants that may vary significantly.

The NEH document is the largest in the corpus, with 1,229 words, and has the lowest vocabulary density. "Data" appears 62 times in the document, and in 9 instances occurs as part of the phrase "data management plan". "Management" appears an additional 11 times in the document (for a total of 20 uses), 8 of which are in the phrase "data management". This indicates a clear emphasis on the importance of this genre of writing – the data management plan. Further, the research practice – data management – which the Executive Summary is introducing to the disciplines the NEH serves, is also emphasized through repetition.

Beyond the specifics of the individual document findings, this textual analysis demonstrated several more general findings to the project team. What guidance is available from governmental agencies concerning the priority and planning of research data management policies and practices varies significantly and is highly influenced by the culture of the specific agency and agency-specific biases. Each of the major U.S. funding agencies examined in the course of this research demonstrates

such agency-specific variations in the requirements and guidance provided to researchers and universities.

Examination of Institutional Policy Responses

In order to better understand the response of research universities to agency requirements for the retention and sharing of research data the project team conducted several types of investigation into institutional policies and responses.

The project team began by identifying a pool of institutions to study. Since the issues of data management will likely be most salient for institutions that receive significant amounts of grant funding involving the production of research data, we used this as a selection principle. Public records of the NIH and NSF were reviewed to identify the most frequent and largest recipients of research awards. A noticeable drop-off in total annual funding occurs after approximately the top two hundred recipients for each agency. Further, there is significant overlap between the lists of the top two hundred awardees of the two agencies. Merging the two lists produced a combined list of approximately 220 entities. This merged list was further winnowed by eliminating entities that were not associated with a university (foundations, museums, etc.), in the interests of studying a group that is roughly comparable. This produced a pool of 212 institutions to examine.

Throughout 2012 using a variety of search techniques (primarily a combination of Google searches and direct email inquiries to campus representatives) the project team scanned for public institutional policies produced at the level of the provost or office of research for this group of 212 institutions. This scan ultimately identified only 38 institution-wide polices concerned data management, or roughly 18% of the institutional pool examined. Examination of these 38 policies showed that many predated the NSF's requirement, and were likely developed, at least in part, in response to the earlier plan requirement of NIH for data management.

Of those institutions lacking publicly available policies governing the retention and sharing of research data, it is possible that some have such policies, but that they are not being made public. It is also possible that some institutions are in the process of revising their data management policies, or drafting new policies, in response to the demands of the NSF and other funding agencies.

Surveys

The project team distributed an online questionnaire, entitled the DataRes Online Survey (DROS). The survey invitation was distributed primarily via listservs, mostly listservs associated with the American Library Association. The survey had conditional branching logic that resulted in between 7 and 15 questions for individuals. There were 231 respondents. 76% of respondents were librarians, with others identifying themselves as researchers or other academic administrative officials.

The survey confirmed findings from the policy scan. Very few individuals (9%) indicated that their institution had a policy governing the retention and sharing of research data (Figure 1). 72% indicated that their institution did not have such a

policy, and an alarming 19% said "I don't know", which could be equated with a "No" response, since the participants' lack of knowledge could suggest that even if a policy was in place, it is not being enforced to a degree that would require awareness or procedural changes.

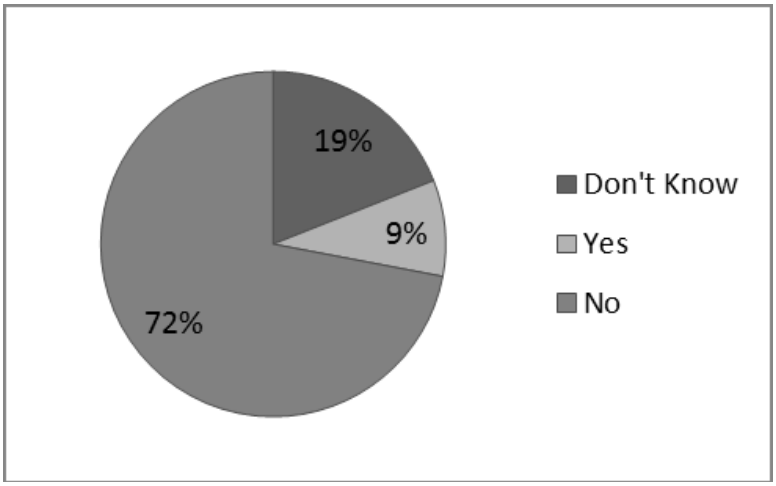

Figure 1: Responses to the question: "Does your institution have a policy governing the retention and sharing of research data?"

Immediately following this question on the existence of a policy, we asked the participants to indicate how strongly they agreed or disagreed with the following statement: "I believe that an institution-wide data management policy is valuable" (Figure 2).

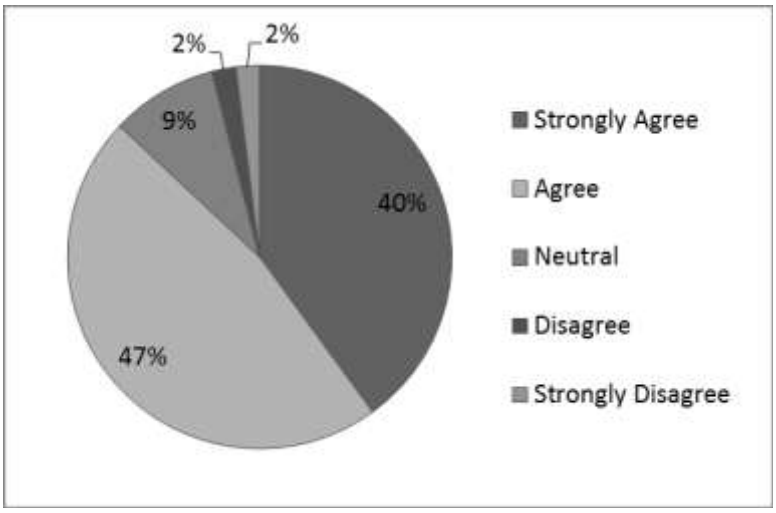

Figure 2: Chart representing participant response to the statement, "I believe that an institution-wide data management policy is valuable."

The majority (87%) indicated either agreement or strong agreement with the statement, while only a combined 4% disagreed or strongly disagreed. The remaining percentage showed a neutral opinion on the subject. These responses suggest that stakeholders are eager to see their institutions make a clear proclamation on the

subject of research data management, which aligns with the responses received in other qualitative inquires conducted by the project team (see the section on focus groups below).

The survey asked additional questions on support and infrastructure in order to get a baseline understanding of how institutions are currently handling these needs. As a starting place, participants were asked where their research data is physically located, and more than half the respondents reported that it was kept on a "Local computer or external hard drive" (54%).

Other questions in the survey confirmed a finding of the 2010 ARL survey, namely that there is a general perception that data management are most effective when managed with a collaborative approach across multiple institutional departments and offices, especially including the library, campus computing centers, and institutional research administration offices.

A second project survey is planned for 2013; the focus of this second survey will be to assess the opinions of academic administration officials such as VPs of Research, Deans, and higher-level administrators in order to get their perspective on these issues. The project team also hopes to gain a better picture of what changes —if any —we can expect to see in the future, based on current planning and priorities.

Focus Groups and Interviews

Several focus groups and individual interviews were conducted in late 2011 and early 2012 as a way of investigating the issues of the project in a more open-ended way. Focus groups were conducted in person, and ranged from five to eight participants, who were invited to share their views on a series of data management questions. Focus groups included a variety of different types of individuals, including program officers from the National Science Foundation and National Science Board, librarians, library administrators, and data scientists. Participants came from a variety of institutions as well, ranging from large research universities to suburban community and teaching colleges. Individual interviews were carried out over telephone or videoconference, and again invited individuals to share their opinions in a series of open-ended responses. The transcripts of these focus groups and interviews affirmed conclusions drawn from the survey, such as the fact that most institutions do not yet have coherent institution-wide policies or collaborative interdepartmental data management programs, but most individuals believe that it is very important to develop such policies and programs.

Both the survey and the focus groups also reinforced another finding hinted at by the 2010 ARL survey, namely that librarians strongly feel that they should play a role in data management efforts, they are poorly equipped as of yet in terms of expertise and resources to provide data management services. In the small number of cases in which institutions are developing data management support services, the individual librarians being charged with participating in these efforts do not feel adequately prepared in terms of training, tools, or funding. This is another troubling finding that will be discussed in the section on conclusions and prospects.

iCAMP Project Synergistic Findings

As mentioned, the iCAMP Project (http://icamp.unt.edu) is another UNT project allied with DataRes. The project entails a three-year effort (2011-2013) to build capacity for educating librarians and researchers for digital curation and data management. This effort will result in an open body of curriculum materials and associated set of graduate level courses offered at the University of North Texas. Another aspect of this capacity building effort is to produce new digital curators and data managers ready for the challenges of digital information curation and preservation and data management.

One of the most synergistic activities in iCAMP was the initial project research that informed the development of course objectives and other curricular materials. In order to develop a curriculum that would be effective in preparing new LIS professionals to work in the emerging roles of data management programs, an analysis was conducted of job advertisements. The methodology was as follows.

Several sources of job advertisements were actively searched and filtered for position postings that matched an extensive set of phrases and keyword variations associated with digital curation and data management. The monitored sources included ALA JobLIST, ARL's Job Announcements, LIS Jobs, and Digital Curation Exchange Jobs; a set of sources that collectively cover all the regions of the United States and Canada. Using the topical filters a group of 110 job advertisements were collected between October 2011 and March 2012. This group of advertisements comprised a textual corpus of entries that were analysed using another textual analysis tool, the NVivo qualitative analysis software. A content analysis coding system was applied to the corpus to categorize the advertisements into different dimensions (position title, educational requirements, experience, skills, knowledge, etc.). The project team then studied the results to identify patterns of specific characteristics and requirements that recurred across the corpus.

This analysis was used to produce a set of competencies that in turn drove the development of curriculum, and the specific expression of this curriculum in the form of four new courses that are now starting to be taught at UNT. What is synergistic about the results of the iCAMP project is that it offers a prescription for addressing the gaps in training identified by many librarians in the DataRes study. However there are evident challenges in implementing data management curricula in LIS programs across the country; notably the question of how quickly LIS programs will themselves be able to grow the expertise and resources necessary to implement new curricula to prepare data management professionals. These and other problematic aspects of our current situation will be discussed in the next section.

Conclusions and Prospects

The DataRes and iCAMP projects have resulted in findings that provide a lens into the changing functions and circumstances of data management in research universities, findings which represent both opportunities and troubling challenges for not only the library profession but the academic research community as a whole.

A troubling finding is the disconnect between beliefs and practice. Most (87%) respondents to the DataRes survey strongly believe in the importance of implementing institution-wide policies for research data management. Yet, despite this widespread belief, and despite the fact that research data management plans have been mandated for more than two years in the United States, there has not been a robust response from universities in terms of institutional policy implementations. An overwhelming 82% of respondents to surveys indicate that their institutions have still not implemented any institutional policy to address institutional research data management needs. And while both survey responses and focus group discussions indicate a willingness by libraries to participate in institutional data management programs, there is also strong reason to believe that libraries currently lack the expertise and resources to effectively contribute to prospective programs.

Both net discussions and the focus groups conducted in the DataRes project suggest that throughout 2012 many universities have been considering implementation of data management policies, or at least tentative first steps to support better data management practices. Yet, the stance of many university administrators remains a wait-and-see attitude. Although there have been repeated claims that the new funding agency data management plan requirements are no longer optional, agency officials themselves are clearly conflicted when interviewed in focus groups. Some believe that strong data management requirements should be enforced for awarded projects, although there is little consensus as to what form those requirements should take, even for disciplines with well-established research data practices. Contrariwise, some officials also take a wait-and-see attitude, believing that consensus as to good data management practices must evolve among researchers and become evident through the results of the peer review process of grant reviews. The countervailing viewpoint is that researchers may have no motivation to self-impose additional requirements that do not directly serve the purposes of academic career advancement. And if the data management plan requirements of grant programs prove to be empty of significance and do not come into play when award decisions are made, then university administrations will feel justified in not investing additional resources in data management programs, since such investments do not result in significantly increase likelihood of research funding.

While the affiliated iCAMP Project has created a prospective curriculum designed specifically to educate new LIS professionals seeking to enter data management roles, and will be making the syllabi and other materials associated with this curriculum freely available to others, the question remains as to whether or not the existing base of LIS faculty throughout the United States will be able to adjust their educational programs and teaching quickly enough to address the rapidly changing state of the field. There is a great deal of inertia in university faculty, and it is an open question as to how quickly LIS programs will be able to adapt to the pace of change in the uncertain landscape of data management.

Finally, if neither libraries nor other institution-wide data management stewards emerge within the next decade, what are the likely prospects for research data? Particular DataRes analysis interviews with scientific instrumentation support personnel suggest that junior members of scientific research teams who are unprepared for data stewardship activities may continue to be called upon to function in the role of data librarians. What is emerging is a potential new class of individuals

who must function as the librarians for the digital age yet do not understand themselves as librarians or have effective training for data librarianship. These individuals, often post-doctoral scholars with temporary appointments as visiting scientists, are being called upon to perform the work of long-term custodians of preservation and access with no formal preparation or training for the data curation roles thrust upon them.

While the DataRes Project has so far been focused on descriptively documenting the responses of universities at the institutional policy level to data management needs, our intent now is to investigate emerging data management practices, and to document to the extent possible whether these emerging practices are effective or dysfunctional. We hope to document effective prescriptions for success. This is an important period of rapid evolution in data management practices, during which both descriptive and prescriptive analysis of the changing landscape is needed.

Acknowledgements

We would like to acknowledge the support of the U.S. Institute of Museum and Library Services which funded this work.

DataRes Project Team: Dr. Martin Halbert (PI), Dr. William Moen (Co-PI), Dr. Spencer Keralis, Shannon Stark.

iCAMP Project Team: Dr. William Moen (PI), Dr. Jeonghyun Kim (Co-PI), Dr. Martin Halbert (Co-PI), Mark Phillips, Ana Krahmer, Brenda Cantu, Paulette Lewis, Jacqueline Salter, Edward Warga, Joseph Helsing.

References

Association of Research Libraries. (2006) To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering. Association of Research Libraries: Washington, DC. Retrieved from http://www.arl.org/bm~doc/digdatarpt.pdf

Association of Research Libraries. (2007) Agenda for Developing E-Science in Research Libraries: ARL Joint Task Force on Library Support for E-Science Final Report & Recommendations. Association of Research Libraries: Washington, DC. Retrieved from http://www.arl.org/bm~doc/ARL_EScience_final.pdf

Berman, F.; et al. (2010) Sustainable Economics for a Digital Planet: Ensuring Long Term Access to Digital Information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. San Diego Supercomputer Center. Retrieved from http://brtf.sdsc.edu/

Hswe, P., & Holt, A. (2010). Guide for Research Libraries: The NSF Data Sharing Policy. Association of Research Libraries. Retrieved from http://www.arl.org/rtl/eresearch/escien/nsf/index.shtml

Monash University. (2009). Data Management Planning. (Website) Monash University: Melbourne, Australia. Retrieved from http://www.researchdata.monash.edu.au/guidelines/planning.html

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

Soehner, C., Steeves, C., & Ward, J. (2010). E-Science and Data Support Services: A Study of ARL Member Institutions. Association of Research Libraries: Washington, DC. 2010. Retrieved from http://www.arl.org/bm~doc/escience_report2010.pdf

Swan, A. & Brown, S. (2008). The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs. Report to the JISC. July 2008. Retrieved from http://ie-repository.jisc.ac.uk/245/

Van den Eynden, V., Corti, L., Woollard, M. & Bishop, L. (2009). Managing and Sharing Data: A Best Practice Guide for Researchers . UK Data Archive, University of Essex: Colchester, UK. Retrieved from http://www.data-archive.ac.uk/media/2894/managingsharing.pdf