Research Data Management

Principles, Practices, and Prospects

November 2013

The Denton Declaration:
An Open Data Manifesto

Council on Library and Information Resources

Research Data Management

Principles, Practices, and Prospects

November 2013

Council on Library and Information Resources

ISBN 978-1-932326-47-5
CLIR Publication No. 160
Published by:

Council on Library and Information Resources
1707 L Street NW, Suite 650
Washington, DC 20036
Web site at http://www.clir.org

Cover art derived from Mariette Papić's poster for The Denton Declaration. Also included are images from Shutterstock.com.

Contents

About the Authors. iv
Acknowledgments vii

Prospects for Research Data Management, *by Martin Halbert* 1

Research Data Management in Policy and Practice: The DataRes Project, *by Spencer D. C. Keralis, Shannon Stark, Martin Halbert, and William E. Moen* 16

The Denton Declaration: An Open Data Manifesto 39

Why, How, and Where We're Going Next: A Multi-Institution Look at Data Management Services, *by Kiyomi Deards*. 43

Responses to Data Management Requirements at the National Scale, *by Chris Jordan, Maria Esteva, David Walling, Tomilsav Urban, and Sivakumar Kulasekaran* 64

Dilemmas of Digital Stewardship: Research Ethics and the Problems of Data Sharing, *by Lori M. Jahnke and Andrew Asher* 80

About the Authors

Andrew Asher is the assessment librarian at Indiana University Bloomington, where he leads the libraries' qualitative and quantitative assessment programs and conducts research on the information practices of students and faculty. Asher's most recent projects have examined how "discovery" search tools influence undergraduates' research processes, and how university researchers manage, use, and preserve their research data. Prior to joining Indiana University, Asher was the digital initiatives coordinator and scholarly communications officer at Bucknell University, where he managed the library's open access and scholarly communication initiatives, including the passage of an institutional open access mandate. An ethnographer and anthropologist by vocation, Asher holds a Ph.D. in sociocultural anthropology from the University of Illinois at Urbana-Champaign.

Kiyomi Deards has been an assistant professor and science librarian with the University Libraries at the University of Nebraska-Lincoln (UNL) since 2010. She is currently the subject librarian for Chemistry, Biochemistry, Physics, and Astronomy. Deards serves on several university committees, the American Library Association's New Members Round Table mentoring committee, the Association of College and Research Libraries 2013 volunteers committee, and UNL Libraries data curation committee. Deards has presented publicly on a variety of topics including succession planning in academic libraries, developing communities of practice, marketing yourself in science, data discoverability, and promoting open access repositories. Her research interests include issues of equity and management in libraries and science.

Maria Esteva is research associate/data archivist in the Visualization and Data Analytics Group at the Texas Advanced Computing Center (TACC), University of Texas at Austin. Her research focuses on digital archives and digital preservation. Her position involves developing scientific data collections, implementing digital archiving and preservation strategies for scientific datasets, and the use of information visualization as a tool for archival processing. Maria earned her Ph.D. in Information Science from The University of Texas at Austin. Prior to joining TACC, Maria was an assistant instructor in the UT Austin School of Information. She also worked as a graduate research assistant in the UT Austin Digital Library Services Division where she worked in a variety of areas including text encoding, metadata, institutional repository implementation, and image collections.

Martin Halbert is dean of libraries and associate professor at the University of North Texas. He also serves as president of the MetaArchive Cooperative, a growing international digital preservation alliance of cultural memory organizations that was one of the founding partners of the U.S. National Digital Preservation Program. He has served as principal investigator for grants and contracts totaling more than $6 million during the past six years, funding more than a dozen large-scale collaborative projects among many educational institutions. His doctoral research and subsequent projects have focused on exploring the future of research library services. He has previously worked for Emory University, Rice University, UT Austin, and the IBM Corporation.

Lori Jahnke is the anthropology librarian and an adjunct lecturer in the Department of Anthropology at Emory University. She holds a Ph.D. in Biological Anthropology and her research interests include the biological impact of colonization and social stratification, human paleopathology, the ancient Andes, modeling human population structure, non-textual systems of information organization and communication, and the sociocultural implications of information economies. Prior to joining the Robert W. Woodruff Library at Emory University, Lori was a CLIR Postdoctoral Fellow at The College of Physicians of Philadelphia and the University of Pennsylvania. Her primary project was to develop the Medical Heritage Library (http://www.medicalheritage.org) as a multi-institutional collaboration for digitization in the health sciences. Lori was a research lead for the Sloan-sponsored CLIR/DLF study on data management practices among university researchers.

Chris Jordan is a research engineering/scientist associate and manager of the Data Management & Collections Group at the Texas Advanced Computing Center (TACC), University of Texas at Austin. He joined TACC in April 2008 to work on issues related to storage and I/O for high-performance computing, and to long-term digital preservation. He is currently responsible for deploying research data infrastructure to serve the University of Texas System campuses and the national community. He has led data infrastructure activities for the National Science Foundation's TeraGrid and XSEDE projects. Before joining TACC, Jordan spent four years at the San Diego Supercomputer Center, where he helped develop and deploy the GPFS-WAN global file system currently in production across multiple TeraGrid resource providers. Jordan has nearly a decade of experience working with some of the largest supercomputers and storage systems in the world, and has worked with a variety of architectures and scientific applications.

Spencer D. C. Keralis is director for digital scholarship and research associate professor with the University of North Texas Libraries' Digital Scholarship Co-Operative. His research has appeared in *Book History*, *American Periodicals*, and the CLIR report *The Problem of Data*. He has held a Mellon Fellowship at the Library Company of Philadelphia, a Legacy Fellowship at the American Antiquarian Society, and served as a CLIR Postdoctoral Fellow in Academic Libraries with the UNT Libraries. His current projects focus on the implications of social media, digital curation, and data management for the future of the humanities.

Sivakumar (Siva) Kulasekaran is a research engineering/scientist associate in the Data Management & Collections Group at the Texas Advanced Computing Center (TACC), University of Texas at Austin. Before coming to TACC in 2012, Kulasekaran was an IT analyst at the Center for Computation & Technology at Louisiana State University (LSU), where he was involved in the development of Stork data placement scheduler and Petashare, a distributed data archival and visualization project. Kulasekaran received his Ph.D. in Computer Science from Mississippi State University in 2009.

William E. Moen is associate dean for research in the College of Information and has been a faculty member at the University of North Texas since 1996. He also serves as director of the Texas Center for Digital Knowledge, and as associate professor in the Department of Library and Information Sciences. Moen's projects explore: (1) the representation of information objects through various metadata schemas; (2) the design and implementation of digital repositories to store information objects and their metadata; (3) flexible approaches for discovery, findability, and retrieval of digital information; and (4) the challenges of digital curation, data management, and data re-use. He is currently principal investigator on an Institute of Museum and Library Services grant project to develop four graduate courses to address digital curation and data management education and training for the emerging workforce.

Shannon Stark recently left the University of North Texas for a job in the corporate field. While at UNT, she was the strategic projects librarian. Her responsibilities ranged from performing research for federal grants—such as the DataRes project—to managing the strategic initiatives of the Libraries. She also was the primary coordinator for library-driven events such as UNT's Annual Open Access Symposium, which is now entering its fifth year.

Tomislav Urban is senior software developer in the Data Management & Collections Group at the Texas Advanced Computing Center (TACC), University of Texas at Austin. In this role, he is responsible for database design and application programming, web programming and technologies, Geographic Information Systems (GIS), and relational database technology. In addition, he is working on ways to deliver geospatial data collections to the research community. Prior to joining TACC in 2002, Urban spent six years in Information Technology consulting and software engineering.

David Walling is a software developer in the Data Management & Collections Group at the Texas Advanced Computing Center (TACC), University of Texas at Austin. He joined TACC as a student intern in May 2003 and has been working full-time for the center since January 2005. Walling develops database-driven applications in support of scientific research.

Acknowledgments

The DataRes project team would like to acknowledge the Institute for Museum and Library Services for its generous support in the form of a Laura Bush 21st Century Librarians Grant which made this research, and this publication, possible. Thanks to Charles Henry and the CLIR team, in particular Kathlin Smith, for producing this publication. We are grateful to our contributors, Lori Jahnke, Andrew Asher, Chris Jordan, and Kiyomi Deards, and their co-authors, for sharing the products of their labor in this volume. Conversations with Katherine Skinner of the Educopia Institute, Jim Mullins of the Purdue Libraries, and Cathy Hartman of the UNT Libraries made significant contributions to the intellectual work of this research. Thank you to the participants in the Denton Declaration Workshop (listed on pages 41–42) whose contributions were essential to framing our thinking on the principles of research data management; we could not have asked for more engaged collaborators. We are grateful to Mariette Papić for designing the manifesto poster. Thanks to Rachel Frick of the DLF, and Joan Lippincott and Cliff Lynch of CNI for allowing us to share our research in your forums and to piggy-back focus groups and other events on your meetings. We are grateful for the time and insights of everyone who participated in our focus groups and corresponded with us on this topic over the course of the project. Thank you to Joan Cheverie of EDUCAUSE for offering us their unique platform to disseminate our work. To Shannon Stark, our colleague who has left the fold, we miss you and wish you well. Ephraim Freese and Monica Ugartechea were undergraduate assistants, and Anjum Najmi, now a CLIR Fellow, was our graduate assistant during this research. We could not have done it without their help. Finally, we wish to acknowledge, in no particular order, the generosity and support of the following individuals who contributed to this research in ways that, though intangible, were nonetheless essential: Barbara Halbert, Jim and Linda Keralis, Elliott Shore, Christa Williford, Lauren Coats, Denise Perry Simmons, Korey Jackson, Laura Waugh, Kris Helge, Daniel Alemneh, Jeonghyun Kim, Tyler Walters, John West, Nathan Hall, Daniel Burgard, Mark Phillips, and Dreanna Belden.

Prospects for Research Data Management

Martin Halbert

The challenge of ensuring long-term preservation of and access to the outputs of scientific research, especially data sets produced by publicly funded research projects, has become a prominent topic in the United States. In 2011, the two-year DataRes Project was initiated at the University of North Texas to document perceptions and responses to this emerging challenge in U.S. higher education and to explore ways in which the library and information science (LIS) profession could best respond to the need for better research data management in universities. This chapter will highlight some of the most provocative findings of the DataRes Project on the topic of research data management in higher education and then consider possible research data management (RDM) scenarios for the future and the implications of these scenarios.

The DataRes Project sought to document and understand a critical developmental moment, when many universities were starting to articulate the conceptual foundations, roles, and responsibilities involved in research data management. The project investigated the perspectives of stakeholders (e.g., researchers, librarians, information technology [IT] professionals, sponsored research offices) throughout the research lifecycle. Because it is still too early to draw definitive conclusions about prospective roles for LIS or other professionals in research data management, the DataRes Project instead sought to document basic quantitative and qualitative information about stakeholder expectations, current institutional policies, and the preparation that information professionals will need as they take on emerging responsibilities in this area. Because the project was funded by a 21st Century Librarians grant from the Institute of Museum and Library Services, our aim was to establish a baseline study of research data management practices that institutions can use in developing new curricula and training. The greatest benefit of

this baseline study may be that it brings to the surface fundamental problems in the emerging landscape of research data management responses and interventions in the United States. Our research suggests that effective institutional responses to meet the challenge of research data management may be slow in coming, but are inevitable in the long term.

Context

The DataRes Project is not the first effort to address the topic of research data management. The National Science Foundation (NSF) funds a great deal of research in the United States, and that research generates large amounts of data. In 2003, NSF issued two reports noting the growing perception of an urgent need to build up the national data management capacity. The report from a 2002 workshop, provocatively entitled *It's About Time* and sponsored by NSF, the Library of Congress, and other organizations, called for a national research initiative to "build a foundation for digital preservation practices that government agencies, cultural institutions, businesses, and others urgently require" (Hedstrom et al. 2003, 26). The 2003 report by Atkins and colleagues, in which they coined the term *cyberinfrastructure* and articulated an agenda for scientific investment based on data-intensive research, also identified the risks of not managing research data over time: "Absent systematic archiving and curation of intermediate research results (as well as the polished and reduced publications), data gathered at great expense will be lost" (Atkins et al. 2003, 11).

These and similar calls in the first years of the twenty-first century led to major collaborative efforts, such as the 10-year National Digital Information Infrastructure and Preservation Program (NDIIPP) undertaken by the Library of Congress in collaboration with NSF and many other organizations to explore and better understand the foundations of the new field of digital preservation (NDIIPP 2010). Research data management has been widely debated and discussed. Many discussions of its importance have taken place at meetings of professional groups concerned with the topic; these discussions culminated in a variety of organizational recommendations and position papers, such as those of the Association of Research Libraries (2006) and the National Academy of Sciences (2009). At the same time, those in business and society more generally were carrying on a discussion of the criticality of so-called "Big Data," reflecting the growing recognition that computing technology in all walks of life is generating and accumulating ever more vast amounts of data that, if managed effectively, can be "used to unlock new sources of economic value, provide fresh insights into science and hold governments to account" (*The Economist* 2010).

Virtually all of these discussions agreed on two themes. First, the vast amounts of data that research organizations are accumulating are valuable in potentially game-changing ways *if the data are effectively managed*, and second, very few (if any) research organizations

are currently prepared or mandated for the effective management of such unprecedented quantities of data. The growing consensus on these two points was almost certainly a factor in NSF's decision to issue a new mandate in 2010 that all research proposals submitted to the agency after January 2011 must include a "data management plan" (NSF 2010). Such a plan is now understood to be essentially a description of how investigators will "share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants" (National Science Foundation 2013).

The NSF mandate was neither unprecedented nor an isolated intervention by one federal agency. The National Institutes of Health (NIH) had implemented the first major mandate of this kind in 2003, requiring researchers to comply with data sharing and data management practices (NIH 2003). Other federal agencies were adopting similar policies at the same time that the NSF was doing so; for example, the National Endowment for the Humanities adopted a requirement for data management plans that explicitly emulated the NSF requirement (National Endowment for the Humanities 2013).

The NSF mandate prompted a new round of discussions across the United States, especially among intermediaries such as librarians and other information professionals who devote special attention to the long-term preservation of and access to scientific research results. This attention was evident, for example, in the large number of presentations in the 2010 meetings of the Coalition of Networked Information (CNI) that mentioned either the NSF mandate or related research data management topics. It appeared that the concerns voiced in the 2003 reports cited earlier regarding the long-term survival of research data were about to be addressed. There was real hope in many of the 2010 discussions that the new federal agency mandates would lead universities and other research institutions to rapidly adopt much stronger research data management practices and policies.

Study of Research Data Management Responses

The two-year DataRes project was conceived amidst growing concern over research data management. The aims of the project were (1) to study and document trends in the data management plans and associated institutional policies of research institutions in response to federal requirements, and (2) to determine how the LIS profession can best respond to emerging needs of research data management in universities. In the course of the study, project personnel asked a variety of questions about the emerging research data management responses in the United States:

- What trends and patterns are observable in the data management plans and associated institutional policies now being implemented at research institutions in response to federal requirements?

- What do key stakeholders in the research community (e.g., researchers, administrative officials, librarians, funding agency officials, research equipment vendors) expect in the long-term management of research data generated in universities? What is the role of information professionals in such efforts?
- What skills, infrastructure, training, and other preparation do professionals charged with data management responsibilities need, based on both expectations of stakeholders and observed trends in data management policies now being implemented?

The detailed findings of this two-year study are provided elsewhere in this volume. There have been some undeniable quick accomplishments of the "low-hanging fruit" variety to give researchers at the local level basic advice on dealing with the new NSF mandates. For example, low-cost local university workshops have been held and tools cooperatively devised to help researchers develop data management plans (Sallans 2012). But what has most impressed the DataRes research team in the course of this work is the range of barriers to effective research data management at scale, at both the local and the national levels. Although virtually all stakeholders acknowledge the importance of effective long-term management of research data, a daunting array of barriers hamper the prospects for effective research data management practices and programs.

Barriers to Research Data Management

The barriers that hinder effective research data management are not intractable, but they are real. They must be fully understood if institutions of higher education in the United States are to make headway in overcoming them.

Lack of Funding

The most frequently identified barrier to effective research data management is lack of funding. The vast majority of stakeholders consulted in the DataRes Project believed that research data management is an important need that should be addressed, but felt that it does not receive funding at the level required to build needed infrastructure and programs. This perception is somewhat equivocal. The DataRes surveys show that *some* funding is being devoted to research data management programs, usually through a combination of sources. But the overall sentiment expressed by most DataRes survey respondents was that this funding is very modest in scale and often takes the form of incidental commitments of time by librarians who are primarily tasked with other duties. With few exceptions, it was perceived that most institutions devote an almost inconsequential amount of their budgets to research data management functions.

Research data management programs still seem to be mostly conceptual and prospective at a time when the competing demands to fund existing programs in academia are legion. DataRes discussions with stakeholders, including researchers, librarians, university

administrators, and NSF program officers, repeatedly came back to questions of how to fund these programs at scale. Researchers do not wish to allocate research funds to activities, such as research data management, that they see as occurring outside the scope of research. Librarians see a clear need for long-term preservation and access to research data, but typically are not funded to undertake such functions. University administrators do not have established frameworks to determine the relative priority of research data management in the ecology of programs for which they are expected to allocate funding. NSF program officers see the importance of research data management (hence, the new mandates for data plans), but they expect that the consensus on the relative allocation of funding in grant programs will emerge from the field, primarily from researchers. Many academic stakeholders who are not themselves researchers expect that the funding for research data management programs will come from research grants, but this approach ignores the predominant perspective of researchers that the purpose of grants is to fund research, not to maintain research outputs.

Until the fundamental issue of funding is resolved, research data management programs will not be created at any useful scale. But funding obviously follows from other preconditions, including the existence of institutional mandates, professional preparation, and organizational structures. Unfortunately, there are major deficiencies in these areas as well.

Lack of Organizational Structures

The organizational structures of academia are slow to change. They are largely based on long accepted notions of the archetypical functional parts of a university: the faculty, the administration, the library, and (most recently) business IT management. Although intramural collaboration between these groups is encouraged to advance the basic academic goals of research and teaching, these functional divisions are still largely understood as organizational silos. Research data management is among the priorities that have emerged in recent years to challenge these organizational boundaries (another is course management systems).

The findings of the DataRes Project support the idea that effective research data management practices will require close working relationships between divisions of the university, sometimes to the point of blurring boundaries in uncomfortable ways. Although hybrid organizational structures may be required for effective research data management, there are as yet no clear models for these structures. Organizational structures exist for many reasons, including accountability, allocation of funds, and comprehensibility by those trying to interact with the organization. In the case of traditional types of research outputs (e.g., published print journal articles), stakeholders have a general understanding of how the longstanding organizational structures of academia are *supposed* to work together (whether or not they actually work well together). The functions entailed in effectively managing digital research data do not fit as

neatly into these traditional organizational divisions, although these roles are starting to blur. Libraries are not classically understood as being the primary point of management for digital information created by scholars; however, libraries are slowly being reconceived in digital terms.

Business IT is usually associated with central institution-wide functions, such as accounting and electronic mail, and is not typically considered to be deeply embedded in the work of university research teams. Nevertheless, IT functions have been a growing aspect of large research laboratories for many years. University offices of research are usually focused on the administrative aspects of applying for, receiving, and managing grant awards, not the research outputs after the grants have been expended. Yet, if federal agencies implement more stringent (read auditable) requirements for long-term preservation and access to research outputs, research offices will feel pressure to interject themselves into these longer-term aspects of research. Academia has only started groping tentatively toward an understanding of what organizational structures will best support long-term research data management; the DataRes findings show that more integrated organizational structures work better than silos. A better shared understanding of the skills and roles of the various actors in the research cycle is needed to breach these silos.

Lack of Professional Preparation

The DataRes Project identified the lack of training, certification, and other types of professional preparation as another basic deficiency in academia's readiness for research data management. This is perhaps not surprising, given that data management is still an emerging area and there is no general understanding of its requirements among the different parts of academia, but it is nevertheless a huge deficiency for effective long-term research data management. Yet, almost no one within the academic community receives systematic professional training and certification in the management of research data. Still worse on a more fundamental level, *virtually no one in academia perceives that they have a professional responsibility or mandate for research data management functions*.

The DataRes research indicates that librarians may be the closest to understanding their role in research data management, but the standard curriculum of library schools does not include preparation for managing large bodies of data. Moreover, most librarians are unsure exactly what re-training is most important for such duties. Most stakeholders (including librarians) also acknowledge that libraries cannot manage research data alone, but are not yet certain what mix of professional skills is most appropriate for cross-organizational teams working on research data management functions. There have been some LIS curriculum development activities for digital curation roles that may be relevant to research data management roles; this issue will be taken up in the section on scenarios for professional preparation.

Lack of Priority among Researchers

A recurrent theme encountered in the DataRes Project was that researchers are rewarded primarily for undertaking new research, not for managing the results of prior research. The main reason that researchers do not request grant funds for research data management is that they seek to maximize the proportion of grants devoted to research proper rather than to functions that they see (understandably) as secondary support operations. The idea that grants will increasingly be judged in terms of the quality of their data management plans is still unproven. Because researchers themselves are typically the primary agents that judge the quality of federal research proposals in peer-reviewed panels, it is unclear whether long-term management of research data will become a priority in designing future research projects.

Lack of Institutional Mandates

Finally, no generally understood institutional mandates exist for managing research data effectively. Producing data in the course of research activities has traditionally been understood as part of the task of researchers. The idea that researchers should share cumulative sets of research data to advance larger research agendas is a relatively new concept that may have developed from the experience of groups that worked together on multiyear, multi-institutional endeavors such as the Human Genome Project. But although projects like the Human Genome Project show that large-scale sharing of research data can produce major data sets of long-term significance, there is no consensus on or established expectation for long-term data management by individual researchers or institutions. This lack of consensus results in a lack of institutional mandates or policies regarding research data management.

The DataRes Project findings show that the vast majority of universities in the United States are not yet implementing research data management policies at the institutional level; it is simply too soon. After studying the current landscape of higher education, we concluded (perhaps unsurprisingly) that policies come only after practices have stabilized and become accepted, and this has not yet happened for research data management. Until there are widely shared expectations about research data management practices, the current situation will continue. Without institutional mandates, research data may or may not be preserved in accessible ways; their systematic management will definitely not be an institutional priority. There are some indications that this may change, and they will be discussed in the section on scenarios for the future.

Current Developments

Federal agencies made several notable announcements about research data management during the two years that the DataRes Project studied the issue. The new "Data Sharing Policy" requirements were put into effect for NSF proposals submitted on or after January

18, 2011 (NSF 2010). On March 29, 2012, six federal grant-making departments and agencies announced more than $200 million in grant opportunities for the so-called "Big Data Research and Development Initiative" (Office of Science and Technology Policy [OSTP] 2012).

The Fair Access to Science and Technology Research Act (FASTR) was introduced in both the Senate and the House in early February 2013. If passed, this legislation will require federal agencies to develop policies that ensure rapid access to the products of federally funded research. Shortly after this legislation was introduced, on February 22, 2013, OSTP Director John Holdren issued a policy memorandum entitled "Increasing Access to the Results of Federally Funded Scientific Research," which includes language very much like that in the FASTR bill (OSTP 2013). The OSTP memorandum "directs each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government. This includes any results published in peer-reviewed scholarly publications that are based on research that directly arises from Federal funds . . ." (OSTP 2013, 2). Agencies were given six months to respond, but as of this writing (mid-September 2013), the agencies to which the memorandum was directed have not issued public responses. Although much of the focus of the FASTR legislation and the OSTP memorandum is on published articles as the main category of research results, the memorandum explicitly states at the beginning that "such results include peer-reviewed publications and digital data."

These announcements suggest that federal officials are paying a great deal of attention to research data management. The policies established by various agencies requiring researchers to submit data management plans as part of their proposals were only the first of several steps to encourage researchers and their institutions to increase their efforts to implement more effective practices for the long-term preservation of and access to research data created through federally funded grants. Most of the university responses noted by the DataRes Project were prompted to some degree by the federal announcements, but they also reflected librarians' genuine concerns that research data are significant academic intellectual assets and parts of the scholarly record in their own regard.

Various research stakeholder groups have issued responses to the February 2013 OSTP memorandum well in advance of the deadline given to agencies. The Association of American Publishers (AAP) put forward a proposal in June 2013 titled the Clearinghouse for the Open Research of the United States (CHORUS; AAP 2013), which suggested that publishers should be the primary entities responsible for the long-term management of research results mandated in the 2013 OSTP memorandum. The CHORUS proposal was greeted with skepticism by some researchers (Eisen 2013; Neylon 2013), who questioned whether publishers would be motivated to preserve publications or make them openly accessible to the public.

A coalition of groups including the Association of Research Libraries, the Association of American Universities, and the Association of Public and Land-Grant Universities issued a draft proposal called the Shared Access Research Ecosystem (SHARE), which emphasizes the role of research universities as long-lived, mission-driven institutions focused on creating, preserving, and disseminating knowledge (Association of Research Libraries 2013). The SHARE proposal "envisions that universities will collaborate with the Federal Government and others to host cross-institutional digital repositories of public access research publications that meet federal requirements for public availability and preservation." Other commentary on the OSTP memorandum noted that PubMed Central already provides many of the features requested, and new repositories may simply duplicate those features (Neylon 2013).

What is noteworthy about responses to the OSTP memorandum from CHORUS, SHARE, and other research stakeholders is that they were not responses from the primary audience of the memorandum, namely, the large federal grant-making agencies. Although a consensus on research data management practices has not emerged by 2013, what clearly *has* changed is that many stakeholder groups are now willing to engage in the public debate about research data management. Somewhat disheartening is that the nature of these discussions has been rather heated at times, with the positions taken resembling battle lines drawn in the sand. The DataRes Project findings highlight the need for cooperation between all stakeholders in the scholarly communication cycle, rather than strategies that emphasize the primacy of any single stakeholder group or cluster of stakeholders. The importance ascribed to research data management, not only by federal officials, but also by all stakeholders in the scholarly communication cycle, is likely to continue increasing.

Scenarios for the Future of Research Data Management

The DataRes Project sought to document basic quantitative and qualitative information about stakeholder expectations, current policies, and needed preparation for information professionals taking on emerging responsibilities in data management. This information forms a baseline for institutions as they plan new research data management infrastructures, services, policies, and training programs. Following are possible scenarios for the future in terms of the deficiencies discussed earlier.

Funding Scenarios

Much of the future progress on research data management programs will depend on the availability of funding. The DataRes survey of administrators indicates that the most common practice now is to fund research data management programs through a mixed revenue stream model in which funds from several sources are combined. If this hybrid funding model continues to be the most common means

of funding RDM programs, then the main question is how much funding overall will be achievable for such programs through a combination of sources. One scenario is that the status quo will continue. The early research data management programs now in place, consisting primarily of advisory services for faculty seeking to write data management plans, do not receive significant dedicated funds. The incidental time commitments of those providing advisory services are not much above the level of administrative "noise" and could continue indefinitely without significantly advancing the status of research data management nationally. If the status quo continues in regard to funding, it seems likely that researchers will continue to manage data (if at all) through informal mechanisms, such as USB drive backups in desk drawers. Different scenarios may occur in which one or more of the sources of funding devoted to research data management increases, but the likelihood that new funds will be allocated to research data management depends to some degree on how the other deficiencies are or are not addressed.

Scenarios for Professional Preparation

In at least nine U.S. LIS programs, new curricula and associated certificate programs have been or are being developed to address the new data curation responsibilities of information professionals (Keralis 2012). The well-known DigCCur curriculum development project at the University of North Carolina at Chapel Hill has carefully examined a range of new competencies needed by information professionals tasked with managing digital collections (Hank et al. 2010). The DigCCur program and data curation certificates at other LIS programs around the United States are now beginning to produce graduates who are entering the field, but at a time when (as the findings of the DataRes Project make clear) the future of research data management programs is very uncertain. The real question for scenario analysis comes back to the relative level of priority and funding that research data management programs will receive on university campuses. Sustaining and refining professional preparation programs will require that libraries and other academic employers hire and reward professionals with these skills.

Many library directors consulted in the course of the DataRes Project hope to create research data management programs that will employ new graduates to manage large corpora of data sets. If the number of these programs does increase significantly and the demand for individuals with these skills continues to expand, there is likely to be a national blossoming of professional curricula and certification programs for data curation. If, instead, a perception spreads that librarians with these skills are not in demand, these professional preparation programs will come to be seen as a passing fad. A scenario in which this might occur would be if libraries are largely bypassed in the landscape of emerging responses to research data management. If other stakeholders in the research landscape (especially the growing body of IT managers specializing in operational support of research laboratories) become the primary actors in

establishing research data management programs, there is likely to be less demand for research data management curricula in LIS programs. There could also be a hybrid scenario in which professionals from other disciplinary fields enroll in certificate programs for data curation established by LIS programs. What will drive the demand for professional preparation programs in data curation is a rise in the perceived priority of research data management functions among researchers and institutional mandates for research data management functions.

Research Data Management Priority Scenarios

For long-term research data management to become a higher priority for researchers, they must see clear benefit to be derived from devoting time, attention, and funds to these purposes. It is easy to understand a status quo scenario in which research data management continues to be seen as a low priority or simply as an activity outside the scope of research proposals, but what might a more progressive scenario look like?

There are at least two ways that data management may be assigned a higher priority in research proposals. One possibility is that universities that have been early adopters of strong research data management practices (e.g., Purdue University, University of California, San Diego) will be able to demonstrate the added value of these services prominently enough for researchers at most other institutions to see a compelling competitive need for such services at their own institutions. When research grants regularly begin to feature requests for funds to support local data management, significant progress will start to occur in research data management program development.

The other possibility is that political pressures will build to the point that federal agencies mandate more robust and specific requirements for long-term preservation and access for data produced by grant-funded research, including explicit guidance on requests for research data management funding in applications. This second scenario provides the clearest path to funding research data management programs on a regular basis in the future, but it is also highly speculative because it would entail federal agencies specifying far more prescriptive guidelines for the use of awarded project funds.

Scenarios for Institutional Responses and Organizational Structures

Research data management programs will become a prominent part of the research landscape when they become an expected part of the institutional organization of most universities. The need for research data management is unlikely to go away and will likely continue to grow more prominent over time given that academia and society in general are rapidly becoming more data-driven. The response to the need for research data management can be primarily *reactive* or primarily *proactive*, and these two tendencies will produce quite different outcomes.

In a scenario in which institutional responses are primarily reactive, universities would grudgingly adhere to the stricter compliance measures required by federal agencies and implement the measures only in response to threatened penalties by federal auditors. Standards for research data management might come to be understood as similar to other required compliance standards of performance mandated by the U.S. Office of Management and Budget (such as standards for financial reporting). Universities might be forced to comply with legal strictures by reluctantly creating research data management programs that meet the letter of the law rather than embracing the intent and promise of effective research data management programs.

In contrast, universities could respond proactively by establishing new cross-divisional (perhaps interinstitutional) organizations charged with a strong mandate to preserve and provide access to research data. These organizations could be funded at a level robust enough to develop effectively scaled infrastructure and services in support of this goal. The leadership of many or most universities in the United States would have to be convinced to make a strong commitment to research data management for this proactive scenario to come about, but it could certainly happen. The vision and leadership of individuals in positions of authority will ultimately drive this scenario (and by extension, most of the other positive scenarios discussed). If leaders embrace the concept of research data management in coming years, a proactive scenario could have far-reaching effects across the entire landscape of higher education and research in the United States. Are there reasons to believe that such a scenario could come about?

Conclusions

The DataRes Project has noted several events that may constitute reasons for cautious optimism about the future of research data management. Politicians and federal agency officials are paying more attention to research data management. Federal agencies will soon be required to respond to the OSTP directive with agency plans "to support increased public access to the results of research funded by the Federal Government" (OSTP 2013, 2). Whatever form these individual agency plans may take, they should be understood as incremental steps in guiding institutions and individual researchers toward better stewardship of research data. The actual responsibility for long-term stewardship of research data will fall upon the institutional actors who are tasked with sustaining the various parts of the research endeavor. Are these institutions responding to this challenge?

The CHORUS and SHARE proposals by stakeholder communities demonstrate that those in the field are taking the research data management challenge seriously and that stakeholder groups are engaging in efforts to find solutions to the problems of research data management. Both of these proposals (as well as suggestions to extend existing services such as PubMed Central) offer realistic

approaches that would significantly improve the overall capacity of researchers to manage their data in the future. Each proposal has distinct pros and cons, and a healthy debate is warranted about the relative advantages of these and other new proposals that will no doubt emerge over time.

There are signs that stakeholder groups are coming together to hold constructive debates and discussions. For example, the Research Data Alliance is an international collaboration of many different research stakeholder groups that are addressing research data management as a grand challenge of the same scale as mapping the human genome (Research Data Alliance 2013). This collaboration is a relatively rapid, grassroots community response to the perceived need for multiple institutions to advance the understanding of research data management. Another promising sign of confluence is a September 2013 announcement jointly made by 25 organizations that archive scientific data calling for the creation of models for sustaining and coordinating research data management activities across subject domain repositories (Inter-university Consortium for Political and Social Research 2013).

Finally, DataRes interviews conducted with university administrators reveal that research data management planning efforts are going on at many universities across the United States. During the two years in which the DataRes Project was conducted, the status of these planning efforts has evolved from conceptual debates about whether research data management is a good idea to more practical and specific discussions of who will undertake what efforts with what resources. Although the specific outlines of these programs are still emerging, the overall prospects for research data management are encouraging. The second decade of the twenty-first century will inevitably be a time when the foundations for long-term research data management practices will be established. The shape, scope, and success of these practices will make up the next stage of this developmental process.

Works Cited

Association of American Publishers. 2013. Understanding CHORUS. Press release, June 5, 2013.

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

Association of Research Libraries. 2013. *Shared Access Research Ecosystem (SHARE)*. Draft document, June 7, 2013. Last accessed September 1, 2013 at http://www.arl.org/storage/documents/publications/share-proposal-07june13.pdf .

Atkins, Daniel, et al. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* Washington, DC: National Science Foundation.

The Economist. 2010. Data, Data Everywhere: Special Report on Information Management. February 25, 2010, Special Issue. Last accessed September 1, 2013 at http://www.economist.com/node/15557443.

Eisen, Michael. 2013. A CHORUS of Boos: Publishers Offer Their "Solution" to Public Access. Blog posting, June 4, 2013. Last accessed September 1, 2013 at http://www.michaeleisen.org/blog/?p=1382.

Hank, Carolyn, Helen Tibbo, and Christopher Lee. 2010. DigCCurr I Final Report, 2006-09: Results and Recommendations from the Digital Curation Curriculum Development Project and the Carolina Digital Curation Fellowship Program. March 1, 2010. Last accessed September 1, 2013 at http://www.ils.unc.edu/digccurr/digccurr_I_final_report_031810.pdf.

Hedstrom, Margaret, et al. 2003. *It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation*. Washington, DC: National Science Foundation and Library of Congress.

Inter-university Consortium for Political and Social Research (ICPSR). 2013. Sustaining Domain Repositories for Digital Data: A Call for Change from an Interdisciplinary Working Group of Domain Repositories. Press release, September 16, 2013. Last accessed September 17, 2013 at http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/announcements/2013/09/sustaining-domain-repositories-for.

Keralis, Spencer. 2012. "Data Curation Education: A Snapshot." In *The Problem of Data*, by Lori Jahnke, Andrew Asher, and Spencer D. C. Keralis. Washington, DC: Council on Library and Information Resources. Available at http://www.clir.org/pubs/reports/pub154.

National Academy of Sciences, Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age.* Washington, DC: National Academies Press. 2009.

National Digital Information Infrastructure and Preservation Program (NDIIPP). 2010. *Preserving Our Digital Heritage: The National Digital Information Infrastructure and Preservation Program 2010 Report. A Collaborative Initiative of The Library of Congress.* Washington, DC: Library of Congress.

National Endowment for the Humanities. 2013. *Data Management Plans for NEH Office of Digital Humanities Proposals and Awards.* National Endowment for the Humanities. Last accessed September 1, 2013 at http://www.neh.gov/files/grants/data_management_plans_2013.pdf

National Institutes of Health. 2003. Final NIH Statement on Sharing Research Data: NOT-OD-03-032. February 26, 2003. Last accessed September 1, 2013 at http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html.

National Science Foundation. 2010. Dissemination and Sharing of Research Results: NSF Data Sharing Policy. Washington, DC: National Science Foundation. Last accessed September 1, 2013 at http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

National Science Foundation. 2013. *Grant Proposal Guide (GPG):* Chapter VI, Other Post Award Requirements and Considerations. Washington, DC: National Science Foundation. Last accessed September 1, 2013 at http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/aag_6.jsp.

Neylon, Cameron. 2013. Chapter, Verse, and CHORUS: A First Pass Critique. Blog posting, June 7, 2013. Last accessed September 1, 2013 at http://cameronneylon.net/blog/chapter-verse-and-chorus-a-first-pass-critique/.

Office of Science and Technology Policy. 2012. Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments. Press release, March 29, 2012. Last accessed September 1, 2013 at http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.

Office of Science and Technology Policy. 2013. Memorandum for the Heads of Executive Departments and Agencies, Subject: Increasing Access to the Results of Federally Funded Scientific Research, February 22, 2013. Last accessed September 1, 2013 at http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Research Data Alliance. Website, "About" page. Last accessed September 1, 2013 at http://rd-alliance.org/about.html.

Sallans, Andrew. 2012. DMP Online and DMPTool: Different Strategies Towards a Shared Goal. In *The International Journal of Digital Curation* 7(2):123–129. Last accessed September 1, 2013 at http://www.ijdc.net/index.php/ijdc/article/view/225.

Research Data Management in Policy and Practice: The DataRes Project

Spencer D. C. Keralis, Shannon Stark, Martin Halbert, and William E. Moen

Abstract

In this paper, we report findings of the DataRes Project, a two-year project funded by the Institute for Museum and Library Services (IMLS). We examine the perceptions of library professionals faced with supporting federal funding agency mandates for research data management plans, describe the state of data management requirements at major federal funding agencies, discuss our findings about the policy landscape at the top National Science Foundation (NSF) and National Institutes of Health (NIH) awardee institutions in the United States, and describe examples of robust responses to the needs of researchers for data management plan support.

Introduction

In October 2010, the National Science Foundation (NSF) announced its intention to require all grant applicants to include a plan for the retention and sharing of research data in their proposals, effective January 18, 2011. Such a plan—"a supplementary document of no more than two pages labeled 'Data Management Plan' … [which] describe[s] how the proposal will conform to NSF policy on the dissemination and sharing of research results"—is to be included with every application for NSF funding, even if the plan is a statement that "no detailed plan is needed" (NSF 2013). Coming as it did amid the so-called Data Deluge, this data management plan requirement—often described by stakeholders as an unfunded mandate—initiated a furor across the academic world, from offices of research to research teams to academic libraries. Research universities across the United States 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 surveys of the field and major white papers 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 research that documents 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.

Funded by an Institute of Museum and Library Services (IMLS) Laura Bush 21st Century Librarians award, the DataRes Project was initiated at the University of North Texas to examine how research institutions responded to the NSF and other agency data management plan requirements in terms of policy and practical support for researchers, and to evaluate what role, if any, academic libraries and the LIS profession should have in supporting researchers' data management needs. The project, named DataRes as a shorthand mnemonic for the broad themes concerning research data that it examined, was a collaboration between the University of North Texas Libraries, the University of North Texas College of Information, and the Council on Library and Information Resources (CLIR).

Our research took place in a landscape that was changing as rapidly as things possibly can at the intersection of two monstrous bureaucracies—the grinding point where the tectonic plates of federal agency and academic administration meet. The most appropriate metaphor for the changes that we observed over the course of our research is neither the antediluvian hyperbole typical of discussions of "big data" nor the glacial or geologic metaphors usually applied to discussions of the academic and the federal bureaucracy. Rather, the changes we observed are evolutionary: slow, incremental change over time, punctuated by radical adaptations to local stimuli. Whether this evolution implies an aspect of survival of the fittest remains to be seen.

Background Survey

The DataRes Project developed in part as a response to a 2010 survey of library professionals at 200 U.S. research institutions. The survey, *Support for Research Data Management among U.S. Academic Institutions*, was an attempt to capture librarians' efforts and attitudes toward the management of research data and to determine the role of librarians in supporting data-intensive research in a digital environment (Moen and Halbert 2012).

To summarize the key findings of that survey as they are applicable to the present discussion, 100 percent of respondents (68 respondents, a response rate of 34 percent) believe librarians should "play a role in managing researchers' digital data." Sorting the

possible roles that librarians may play in research data management into broad categories showed that a strong majority[1] of respondents believe that they should participate in the following aspects of managing data:

- informational (directing scholars to resources that will help them manage their own research data)
- instructional (providing training in the tools and information necessary for curating research data)
- infrastructural (providing space and resources for storing and accessing research data)
- cooperative (making tools and other resources available for scholars' use in managing research data)
- collaborative (actively participating in and guiding scholars' research data management)
- archival (preserving and providing access to research data once a scholar or research project no longer resides at the university)

A general concern of respondents, however, was the necessity of top-down institutional support, including financial support and adequate staff, to meet the needs of researchers in any of those roles. The following is a typical response:

> While probably all of these [roles] are critical in terms of their usefulness to researchers, librarians would not be able to provide these services without substantive institutional support, so I have answered framed by the support for these services.

Other respondents cited "woeful budgetary times" to explain their libraries' inability to provide data management support, although they acknowledged that such support is critical to the needs of researchers.

In terms of policy, respondents overwhelmingly responded in favor of an institution-wide research data management policy, with 78.2 percent of respondents describing such a policy as "very useful" (39.1 percent, 25 respondents) or "critical" (39.1 percent, 25 respondents). This finding led us to look closely at the policy landscape at top U.S. research institutions and to examine the ways in which libraries have responded so far to the data management needs of researchers.

Agency Guidance Documents

Our research began in July of 2011 with an environmental scan of the guidance for retaining and sharing research data at both the funding agency and institutional levels. We conducted focus groups at conferences and professional meetings with stakeholders in the

[1] The "strong majority" was nearly unanimous; the only responses of "negligible importance" were in the fields of cooperation (2 responses), collaboration (3 responses), and infrastructure (1 response).

research data management process.[2] We identified and compared the responses of academic libraries to the data management needs of researchers.

Among the federal funding agencies, only NSF, the National Institutes of Health (NIH), the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and the National Endowment for the Humanities Office of Digital Humanities (NEH–ODH) have policies requiring plans for the retention and sharing of research data (Tufts University 2013). IMLS includes a questionnaire on the management of digital research products in its applications, but does not have a requirement for the retention and sharing of research data (IMLS n.d.).

One challenge in navigating the interagency landscape of data management requirements is that each agency maintains its own standards and formulas for grant applications. There is no consistency across agencies in data management guidance documents or even in general guidance for grant applications. As such, the documents we were able to examine from each representative agency varied, ranging from NSF's *Award and Administration Guide* (2011) to the Final NIH Statement on Sharing Research Data (2003). It is also difficult to find the authoritative document on a particular agency's policy. At NSF, each directorate and even individual program solicitations have specific requirements for the data management plan, and in a peculiar bit of circularity, the NSF policy on dissemination and sharing of research data referenced in the *Grant Proposal Guide* refers back to the *Grant Proposal Guide* "for full policy implementation" (NSF n.d., 2013).

As an heuristic exercise, we extracted the text of the data management plan guidance documents from NSF, NIH, and NEH–ODH and entered them into Wordles.[3] The resulting word clouds tell a particular—and surprising—story about the priorities of each agency, and the thinking behind their policies.

The Wordle word clouds suggested that further analysis based on text mining could be fruitful. Text mining, or "distance reading," is a method of quantitative analysis of textual evidence, derived in part from the work of Franco Moretti and other scholars at the Stanford University Literary Lab (Moretti, 2011). Distance reading can make it possible to visualize patterns within texts or networks of associations among a corpus of texts that may be difficult or at least extremely time-consuming to see via close reading of individual texts.

---

[2] Focus groups were held on the following dates and locations:
- December 12, 2011, Washington, D.C. (at the Coalition for Networked Information winter meeting)
- January 20, 2012, Dallas, Texas (between the Association of Library and Information Science Educators and American Library Association midwinter conferences)
- June 27 and 28, 2013, Chicago, Illinois (during the American Library Association annual conference)

[3] "Wordle is a toy for generating 'word clouds' from text that you provide. The clouds give greater prominence to words that appear more frequently in the source text." See http://www.wordle.net/. We limited the clouds to the top 100 words in each document and excluded commonly used words such as articles and prepositions.

Google's N-gram Viewer, for example, can search a vast corpus of texts across a long time period to identify trends in language usage.

More elaborate analytic tools such as Voyant,[4] a suite of tools for lexical analysis developed by Hermeneuti.ca (n.d.), can expand the emphasis-through-frequency data shown in the word clouds to indicate word association, vocabulary density, and word count for individual documents, as well as peaks and trends in frequency and distinctive words in individual texts within a corpus.

Because of this robust suite of analytic tools, we used Voyant to analyze the data management plan guidance documents from NIH, NSF, and NEH–ODH. We applied a Taporware stop words filter provided by Voyant to eliminate commonly used words, such as conjunctions and articles.

National Institutes of Health

In the word cloud for the Final NIH Statement on Data Sharing (figure 1), "data" and "sharing" are prominent (NIH 2003). The NIH policy was instituted in 2003 in a research community already accustomed to strict guidelines for the management of their data (e.g., the Health Insurance Portability and Accountability Act [HIPAA] of 1996). Based on the emphasis illustrated in the word cloud, the NIH policy seems to indicate an agency culture that prioritizes access to research data within the research community served by the agency. "Public" is not prioritized—this is not "open data"—and data sharing is intended to be among researchers.

Fig. 1. Wordle of the Final NIH Statement on Sharing Research Data

The Final NIH Statement on Sharing Research Data contains 869 words. The most frequently used words in the document are "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 that the NIH requirement seeks to foster. The frequency of the agency abbreviation "NIH" (16 uses) underscores the agency's authority as an arbiter of research data practice in the community that it both serves and oversees.

[4] See http://voyant-tools.org/.

National Science Foundation

In contrast, the NSF's *Award and Administration Guide*, Chapter VI.D.4 (figure 2) prioritizes "expected" and "Investigators," but interestingly, the name of the agency is far and away the most prominent item in the word cloud (NSF 2011). This may indicate that the most important thing for the NSF was the mandate itself, not specifically the cultural implications (i.e., the benefit to the disciplines of such a mandate) or the practical implementation of the requirement.

Fig. 2. Wordle of the National Science Foundation's Award and Administration Guide. *Chapter VI.D.4*

With only 350 words, NSF's guidance to researchers is the smallest of the documents; it is, in fact, a small component of a larger document. However, it has the greatest vocabulary density (i.e., the greatest instance of unique words). "NSF" appears seven times in the document; "investigators," five times; and "grantees," "dissemination," and "results," four times each. There is no preponderance of usage of any of the key terms ("data," "management," or "sharing") as in the other agency guidance documents. "Data" appears only three times. "NSF" occurs three times and is paired directly with "grants." The focus, such as it is, appears to be on the authority of the granting agency. Interestingly, each directorate within the NSF gives supplemental guidance for applicants. Further analysis of these documents may be valuable for understanding the distinct ways in which these directorates are soliciting and evaluating data management plans.

At a focus group in December of 2011 with NSF program staff, the National Science Board, and research library administrators, NSF staff clearly articulated the importance of innovation in the disciplines' response to data management plan requirements. Although this approach accounts to some degree for the emphasis of the NSF policy, as well as the various directorate level instructions for plan development, researchers and library professionals at subsequent focus groups have offered other explanations, including a perceived unwillingness on the part of NSF or NSF peer review panels to make funding for data management support and repository services a part

of awards. This perception is contrary to the NSF's stated guidance to researchers that they should include costs for research data management in grant applications, and it derives largely from anecdotal information and library staff understanding of faculty priorities. The received wisdom among focus group participants was that faculty are simply unwilling to include these costs in their grant applications, and this filters up into peer review panels.

National Endowment for the Humanities

It is interesting to contrast NSF's policy with the word cloud generated from the executive summary of NEH–ODH's *Data Management Plans for NEH Office of Digital Humanities Proposals and Awards* (figure 3; NEH). In the NEH–ODH word cloud, "data" is extremely prominent, while "plan" is next in size, and "management" and "NEH" are roughly equivalent. For an agency serving disciplines that are largely perceived as not data-intensive, the focus on planning for the retention and sharing of research data is striking and indicates a shift in disciplinary priorities driven by the digital humanities.

Fig. 3. Wordle of the data management policy of the Office of Digital Humanities, National Endowment for the Humanities

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; in 9 instances, it 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" on its own (as opposed to in the phrase "data management plan." The frequency with which the plan itself is mentioned—9 out of 20 uses—indicates a clear emphasis on the importance of the data management plan. Further, it emphasizes through repetition the research practice—data management—that the executive summary is introducing to the disciplines served by NEH.

Institutional Policies

In July 2011, the authors began an examination of published policies at the provost and office of research levels at the top NSF and NIH awardee schools. To develop the list of research universities for the institutional policy scan, we used the internal reporting tools from NSF[5] and NIH. We selected these agencies because (1) our research agenda was a response to the NSF's requirement for research data management plans, and (2) NIH has the longest standing requirement for research data management plans. We set search parameters for the top dollar awardees for fiscal year (FY) 2010 and extracted the top 200 awardees from each agency. We synchronized the resulting reports, removing duplicates, stand-alone institutes, and individual awardees. The resulting list of 197 institutions constituted the pool for the policy scan.

We excluded IMLS because, while the agency does provide search capabilities for award information,[6] it does not have an effective tool for extracting reports on awardees. Further, it does not offer explicit guidance on data management plans for applicants. IMLS requires applicants to complete a questionnaire, Specifications for Projects that Develop Digital Projects, about data practices, but does not require a data management plan per se on the model required by NSF, NIH, or NEH–ODH. Although we include the NEH–ODH guidance documents in our analysis for comparison purposes, because of the limited budget and scope of the Office of Digital Humanities, we have excluded those awardee schools from our scan (recognizing that many of those awardees may be captured in our report, regardless).

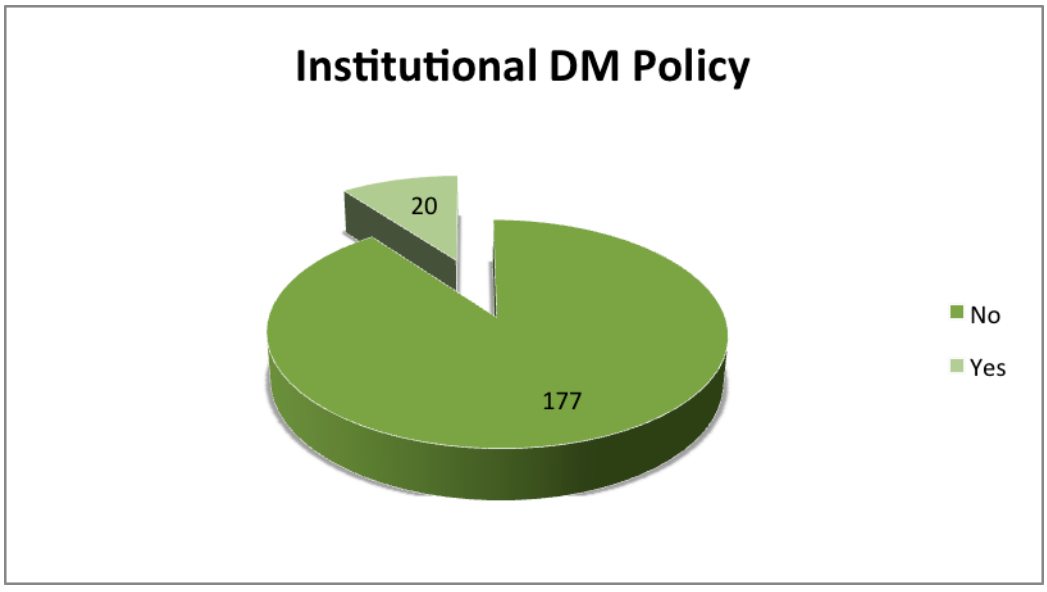

Figure 4: Count of institutions with data management policies

[5] See http://nsf.gov/awardsearch/.

[6] See http://www.imls.gov/recipients/grantsearch.aspx.

By performing Google searches using the institution names, "data management," and "policy" as keywords, then duplicating the search using the institutions' internal site search engines, we determined that only 18 percent (20 institutions) have publicly available policies requiring the retention and sharing of research data; the significant majority (82 percent) did not (figure 4). Many of the existing policies predate NSF's requirement and were likely developed, at least in part, in response to NIH's data management plan requirement. Institutions lacking a policy governing the retention and sharing of research data received in excess of $13 billion in federal research funding from NSF and NIH in FY 2010–2011, a sizable investment of taxpayer money (Table 1).

Policy Found?	Sum of NSF	Sum of NIH	Sum of Total $ Awarded
No	$3,648,260,975.00	$9,653,827,431.00	$13,302,088,406.00
Yes	$802,440,563.00	$3,050,553,480.00	$3,852,994,043.00
Grand Total	$4,450,701,538.00	$12,704,380,911.00	$17,155,082,449.00

Table 1. Funds awarded to research institutions by the National Science Foundation (NSF) and the National Institutes of Health (NIH) in FY 2010–2011

University data retention policies tend to be fairly toothless, using statements of "recognition" of the importance of retaining and sharing research data or "encouragement" for researchers to share data rather than solid institutional mandates. One example is the policy of the University of New Hampshire (UNH), which states, "The University recognizes the importance of data sharing in the advancement of knowledge and education." UNH goes on to restrict sharing of research data "only by specific agreement with persons or entities outside the University except where mandated by Federal funding agencies," (UNH, 2012) further weakening the force of the policy. An index of known policies is available at http://datamanagement.unt.edu/findings.

Focus group respondents overwhelmingly supported the notion that if agency mandates are to be effectively implemented, institutional policy at funded universities will have to fall in line with agency priorities. There are myriad obstacles to this happening, but the greatest are institutional inertia and the liminal status of data as distinct research products. Focus group respondents from institutions with data management policies on the books reported that it can take as long as a decade to establish a provost-level policy. Respondents also described a state of affairs at many institutions in which offices of research are reluctant to engage with the policy or invest in the infrastructure necessary at least in part, for two reasons. First, they perceive the interest in data management as just a trend that agencies are not particularly serious about, and second, the return on investment is difficult to calculate. Offices of research would rather wait until agencies issue a more solid mandate than invest in data services and infrastructure now.

Further, the existing institutional policies are weak, and compliance tends to be limited because the only way to compel faculty to adhere to such a policy is to make compliance a mandate for tenure and promotion, a step no institution is willing to take. Focus group respondents uniformly reported that researchers tend to be reluctant to share data, considering them either residual products of their research or something so idiosyncratic, specialized, or proprietary that they simply prefer not to share the data. Further, data as such are neither valued nor rewarded as research products for tenure and promotion, so they will not be a priority for research faculty whose efforts are focused on publication and the next grant application.

Of those institutions lacking publicly available policies for data management, it is possible that some have such policies, but that they are not public-facing. 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 NSF and other funding agencies. However, given the pace of change at most institutions, it may be years before new policies are implemented.

Data Management in Libraries

Unsurprisingly, given the emphasis of federal agencies on the data management plan itself, many efforts at both the library and institutional levels have focused on support for researchers writing their plans rather than on implementing the plans. For example, as of this writing, more than 100 institutions are registered with the DMPTool, meaning that they have Shibboleth login access to the tools for local researchers to develop plans, as opposed to the eight contributing institutions working on development of the tool (California Digital Library, 2013a). Although this focus is certainly important and reflects the short-term needs of researchers, it does not address what is necessary to implement a data management plan. Development of resources for long-term preservation and access to research data has been uneven and is generally less robust than support services for plan development.

In the course of our research, we identified 32 universities where libraries are providing some level of data management plan support for researchers, but this number is far from comprehensive. Models of support vary widely, from simple web pages linking back to the policy and guidance documents of federal funding agencies, to programs that offer workshops and other practical support for researchers, to infrastructure projects costing millions of dollars per year, or a combination of these. At the University of Minnesota, for example, the libraries provide a range of data management support functions (University of Minnesota 2011). Library specialists can help draft data management plans, consult on funding agency requirements, confer on subject-specific data repositories, and give access to on-campus research computing resources. In collaboration with the university office of research, the library also offers data management workshops to graduate students, faculty, and researchers. These

workshops provide continuing education credit, a requirement for principal investigators. As of August 2010, 250 faculty members had participated in the workshops and consultations, and six departments had invited librarians to conduct workshops for their entire staff (Kelley 2011).

To highlight the diversity of responses to the demand for data management services, the DataRes team organized a panel titled Meeting the Challenge of Data-Management Support in Academic Libraries, for the EDUCAUSE conference in November 2012 in Denver, Colorado. The panel featured Michael Witt, assistant professor of library science at Purdue University; Deb Morley, head of specialized content and services at the Massachusetts Institute of Technology (MIT) Libraries; Sarah C. Williams, life sciences data services librarian at the University of Illinois at Urbana-Champaign; and Ardys Kozbial, chair of the Data Curation Working Group at the University of California–San Diego Libraries. Panelists discussed their libraries' interventions in data management support for researchers at their institutions. The panelists' presentations reflected the findings of the DataRes study, indicating diverse responses to data management ranging from robust, infrastructure-driven models to ad hoc support provided by individual librarians, depending on the resources and culture of a given institution. The panel was broadcast live from Denver as part of the EDUCAUSE online conference, potentially reaching an audience of thousands of participants.[7]

In the libraries at the University of Illinois at Urbana-Champaign, the liaison librarians serve the disciplines most affected by the NSF mandate, and they drive support for data management. The life sciences librarian developed a links web page to give researchers access to information from funding agencies, information about data repositories, and a list of services, including help in developing a data duration profile for research projects using Purdue's Data Curation Profile Toolkit (originally developed in partnership with the UIUC Graduate School of Library and Information Science) (Purdue University n.d.). The library support page includes a link to the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), the institutional repository, which treats data sets as digital objects, but is not explicitly a data repository (University of Illinois at Urbana-Champaign n.d.).

The MIT Library publishes self-help resources for researchers, including checklists for data management plans, and advice on metadata, file formats, and data security. Specialist librarians also provide consulting services for researchers to help with the development of data management plans and the preparation of data for subject-based and institutional repositories (MIT n.d.b). The data management resources page links to DSpace@MIT, the university's institutional open access repository, which is a service of the libraries. DSpace@MIT is described as "stable, long-term storage for their

[7] Details on the panel, including presentations, may be found at http://www.educause.edu/annual-conference/2012/meeting-challenge-data-management-support-academic-libraries.

digital research and teaching output and to maximize exposure of their content to a world audience" (MIT n.d.a). Although it is not exclusively a data repository, DSpace@MIT does support data sets.

The library at the University of California at San Diego offers a suite of data management services under the umbrella of Research Data Curation Services (University of California at San Diego 2013b). Data curation can best be understood as a life cycle approach to research data management that includes planning, the research process itself, preservation and access, and reuse or deaccession. Assistance is provided for the development of data management plans through individual consultations. The library supports long-term preservation in collaboration with the university's Research Cyberinfrastructure (University of California at San Diego 2013a) and the University of California Curation Center, part of the California Digital Library (2013c). The EZID service of the California Digital Library supports digital object identifiers for data and other digital content, thus allowing researchers to create identifiers and to assign and store citation metadata for digital objects (California Digital Library 2013b).

Purdue University offers a range of services centered on the Purdue University Research Repository (PURR). A platform for life cycle data management, PURR provides data management plan development with boilerplate language, collaboration space for projects, digital object identifier service for data sets, and long-term preservation and access to data sets (Purdue University 2013). The suite of services is free to Purdue faculty and graduate students, with nominal costs for projects requiring storage above a standard set of space thresholds. Technologically advanced and infrastructure-intensive, PURR is an exemplary suite of services; however, an institution lacking Purdue's financial resources would find it impossible to replicate.

As we shall see in the analysis of our two surveys, most of the funding for research data management support is coming from libraries themselves, with little or no financial support from offices of research, indirect funds, or other university sources. At one focus group, a visibly irritated program officer declared, "I don't know what you all [librarians] are complaining about. We're sending you business." But this "business" is not the sort that pays. Faculty are accustomed to free library services; in a time when library resources are diminishing, any additional services to meet the needs of research data management requires cuts in traditional library services, such as subscriptions, book purchases, and student services.

Even at the institutions that have been noted, the commitment of the libraries to research data management varies widely. In most cases, either intra-institutional or extra-institutional collaboration has been key to providing research data management support. The level of support that libraries offer is largely contingent on the commitment of both the library and the university administration to provide financial and staff resources. In some instances, a simple links web page is the only intervention sanctioned by university or library administrations unwilling or unable to invest in the infrastructure necessary for long-term preservation and access to research data.

Primary Survey

To identify the current trends in research data management at research institutions, we distributed an online questionnaire, titled the DataRes Online Survey (DROS), in an early stage of the study. The data collected supplies substantial evidence to support the previous findings from the policy scan and corroborates testimonials from the focus groups. Participant responses also influenced the development of a secondary survey that was distributed a year later. Those results are reported in the next section.

The policy scan had indicated a significant lack of institution-wide policies in the top awardee research institutions, and this finding was loosely supported by the DROS (n=231), in which we asked various stakeholders if their institution had a policy governing the retention and sharing of research data (figure 5). To clarify the term *stakeholders*, participants defined themselves as librarians, research faculty, archivists, data managers, deans, and students, among several other professional titles. The survey instrument can be found at http://datamanagement.unt.edu/findings.

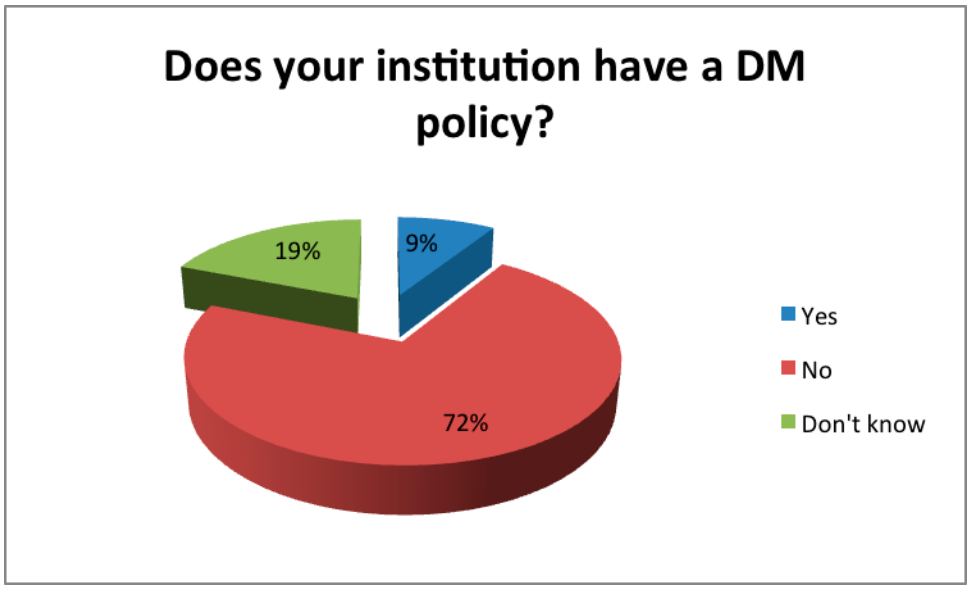

Fig. 5. Participant response to the question, "Does your institution have a policy governing the retention and sharing of research data?"

Only 9 percent of participants answered "yes," while the majority (72 percent) reportedly were employed by or enrolled at an institution that does not currently have a policy. An alarming 19 percent claimed, "I don't know," which could be equated with a "no" response, because the participants' lack of knowledge could suggest that even if a policy were in place, it is not being enforced to a degree that requires awareness or procedural changes.

We applied no mechanism in the survey to prevent multiple individuals from single institutions to respond. Consequently, we expect the percentages for this particular query to reflect higher percentages than those that actually exist. We felt that because most of

the questions pertained to individual preferences and experiences, such a limitation would have hindered our results more than it would have helped overall.

Immediately following this question on the presence or absence 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 6).

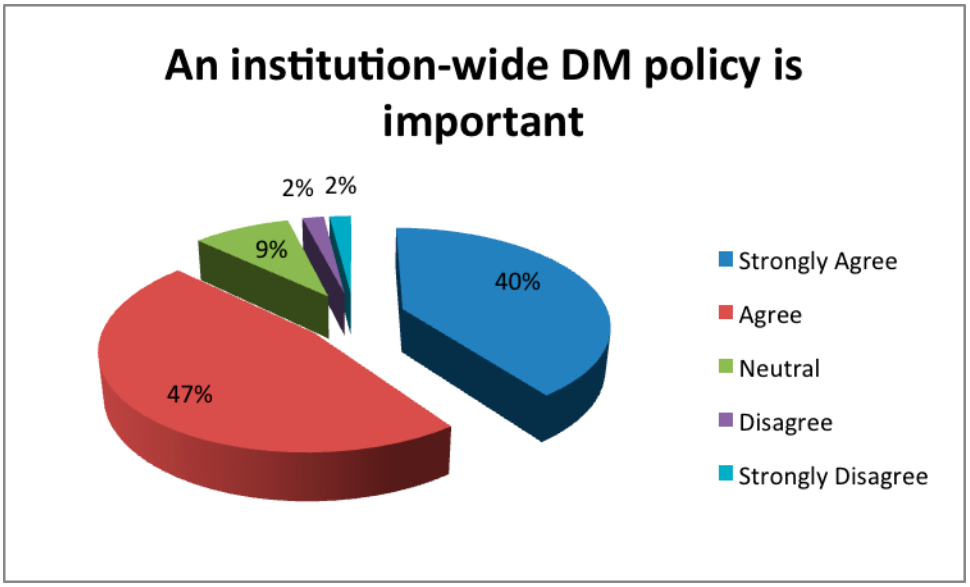

Fig. 6. Participant response to the statement, "I believe that an institution-wide data management policy is valuable."

The majority (87 percent) indicated either agreement or strong agreement with the statement, while only a combined 4 percent 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 is consistent with the responses that we have received from focus group participants. It also invites a more complex conversation on policy enforcement, support, and the infrastructure required for retention and sharing of data. In this initial survey, we focused our inquiries on support and infrastructure to establish a baseline understanding of how institutions currently handle these needs.

As a starting place, we asked participants where their data are physically located, and more than half the respondents reported that data were kept on a "local computer or external hard drive" (54 percent). On a follow-up question, 72 percent said that they would use institutional repository services if they were offered. Table 2 gives more detailed information on the desired data management services, breaking down specific needs and indicating the departments that the respondents believe should be responsible for providing aid.

With the exception of "data storage infrastructure," which was viewed as a responsibility of the information technology services

department, the majority of the participants indicated a preference for repository services to be provided by the office of research or the library. Table 2 is particularly interesting because of the implications for collaboration among departments. The spread of responses suggests that researchers view different aspects of data management as falling under different offices' expertise and that a collaborative approach across multiple departments and offices may be the best way to provide the most desired services.

#	Question	Office of Research	Library	Your Department	Schools of Library and Information Science	School of Computing	Information Technology Services Department	External Service Provider	Responses
1	Workshops on best practices for data management	58	85	13	7	5	26	12	206
2	Workshops on preparing data management plans for funding agencies	74	70	12	7	2	9	7	181
3	Templates for data management plans for funding agencies	73	68	12	5	2	15	9	184
4	Assistance composing data management plans for grant proposals	75	71	14	5	3	7	7	182
5	Data storage infrastructure	22	63	13	5	6	79	15	203
6	Other data management services provided by your institution	18	34	9	4	2	17	9	93

Table 2. Participant response to the question, "If your institution offered the following services and resources, would you take advantage of them? If so, please indicate in which department or office you believe these services and resources should be based. If you don't believe you would use the service, please leave that row blank."

Secondary Survey

To delve deeper and address gaps identified in the first survey, we developed a secondary survey to be sent to vice presidents of research, deans, and higher-level administrators. Although we felt that the first survey had more than addressed the perspective of librarians, we were dissatisfied with the response rate from individuals in administrative positions. Because people in these positions would drive any policy change, we felt it necessary to target them specifically with the DataRes Administrator Online Survey (DRAOS).

We also hoped to gain a better picture of what changes, if any, we could expect to see in the future through the administrators' reports of current planning and priorities. To administer the survey, we assembled a list of 400 contacts at the institutions from the prior sample and e-mailed them directly with the survey link. For 45 days, the link was left active, and responses were accepted. At the end of that period, 33 complete survey responses were collected. Figure 7 shows the makeup of our sample, according to the way in which the individual respondents described themselves.

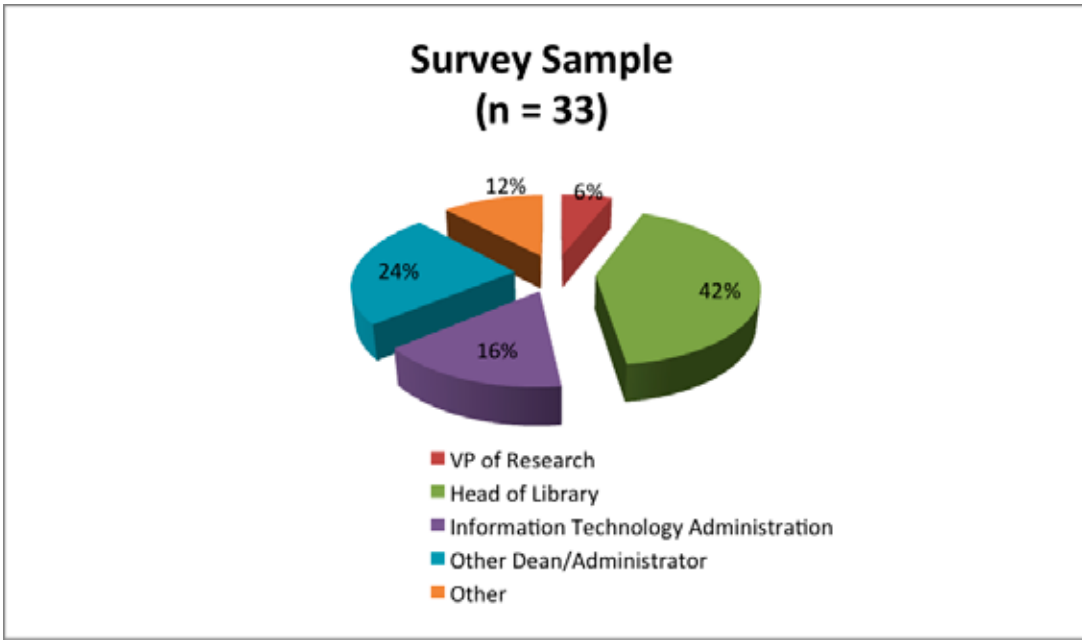

Fig. 7. Administrator responses to the question, "Which of the following best describes you? You may choose more than one."

University librarians, deans of libraries, and library directors made up the majority of our respondents, which we grouped together in the more general category of "Head of Library." The second largest group, "Other Dean/Administrator," included assistant deans and directors from university libraries and graduate schools. Those who defined themselves as "other" were primarily librarians; this group made up only 12 percent of the group, and these titles were often secondary as we allowed individuals to select more than one descriptor, should it apply.

We repeated in the secondary survey several of the questions from our initial survey to compare the librarian versus administrative perspectives. For example, we asked in the first survey which services and resources, if offered, the respondent would take advantage of, and most of our answer choices revolved around the early stages of data management and plan development. When addressing the administrators, we asked them, "At which stage of the data management process do you feel the average researcher at your institution needs the most support?" Figure 8 illustrates the administrators' responses.

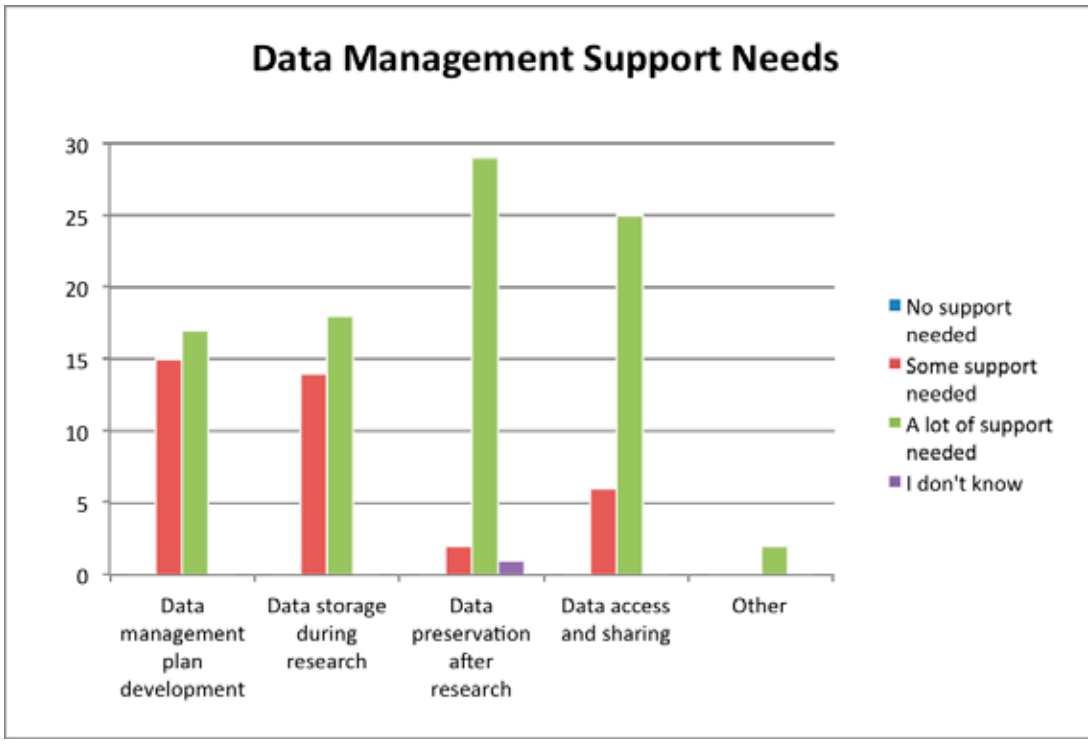

Fig. 8. Administrator responses to the question, "At which stage of the data management process do you feel the average researcher at your institution needs the most support?"

In general, administrators indicated that their researchers needed support at all stages, but they emphasized support during the later stages of "data preservation after research" and "data access and sharing." The two written responses for "Other" for which "a lot of support" is needed were "personal archiving" and "data citation." No responses indicated "No support needed" for any service. Recalling our responses to the DROS, slightly greater numbers indicated a preference for help in the form of "workshops on best practices for data management" and "data storage infrastructure." Meanwhile, fewer people indicated an interest in data plan creation and templates. Not only are the results of the DROS and DRAOS consistent with each other, showing that administrators and librarians at least have similar perspectives, but they also contradict some current practices in the field, which place high emphasis on helping researchers at the earliest stages with data management plans.

In exploring this contradiction, it is important to keep in mind that helping with data management plans and providing workshops for plan development are possibly the easiest and least costly response to the new expectations for data management from the federal granting agencies. Preservation, access, and sharing take additional infrastructure, expertise, and quite a bit more effort and cooperation between researchers and service providers. Answering the questions of how to store data long-term, where to put it, and how long to keep it are far more difficult than organizing hour-long workshops on what a data management plan is. Furthermore, because they were the low-hanging fruit, more early-stage services have

already been created and offered at research institutions, so now it is the later stages that are matters of concern. Whatever the explanation, both of our surveys suggest that late stage data management infrastructure and education are in high demand.

To determine a realistic baseline for the current responses of research institutions to data management needs, we asked in the DRAOS for respondents to report on what services their institutions currently provided and where those services were primarily based (table 3).

#	Question	Office of Research	Library	Individual Departments	Schools of Library and Information Science	School of Computing	Information Technology Services Department	External Service Provider	Total Responses
1	Workshops on best practices for data management	9	19	6	2	0	3	1	40
2	Data storage infrastructure	4	17	15	1	2	25	6	70
3	Continuing education courses on managing research data	1	3	3	4	0	1	0	12
4	Institutional repository for research	3	25	1	0	0	6	1	36
5	Administrative staffing to provide guidance and information regarding research data management	15	23	6	0	0	10	0	54
6	Other	6	8	2	0	0	4	0	20

Table 3. Administrator responses to the question, "What services, if any, does your institution currently offer to manage the retention and sharing of research data? Please indicate all services offered at your institution, and which departments oversee those programs. If a service is provided by more than one department, please indicate all departments involved."

The results showed an encouraging amount of overlap. The most offered service was data storage infrastructure, which was the second most desired service, with only a very small gap between it and the first most desired service (3 votes; see table 3). Also, the administrators reported that such infrastructure is housed primarily in the information technology services department, which was the DROS respondents' preference. For the most desired service, "workshops on best practices for data management," respondents indicated that libraries are the largest single providers of that service, but there were fewer overall offerings available for these workshops than other desirable services. (Note: DRAOS also does not account for multiple individuals responding from single institutions.) This is

a noticeable gap that could be a starting place for institutions hoping to begin providing aid to researchers in this area or for those already providing some services and looking for opportunities to do more.

The spread of services across various departments also echoes the DROS. The library is portrayed as a primary setting for all services except "data storage infrastructure" and "continuing education courses on managing research data," but a notable number of respondents reported many services housed in the office of research and other departments as well. The education courses show up quite evenly distributed between the library, individual departments, and schools of library and information sciences, while the storage infrastructure is housed mostly by the information technology department. Again, this demonstrates an opportunity and need for collaboration across campus, which is currently being fulfilled at some of these institutions.

Next, we wanted to explore how administrators were handling the financial aspect of these new mandates. In the DROS, we had asked, "Do you typically allocate financial resources in your grant proposal budgets for data management?" Only 24 of the DROS respondents (13.95 percent) said "yes" to this question. Curious for more detailed information, we asked the administrators, "Does your institution allocate financial resources for data management?" Then we immediately proposed the follow-up, "From which sources are these funds drawn?" The results appear in figures 9 and 10, respectively.

Only half of the respondents were from institutions that offered financial resources for data management, and of that half, only 17 percent reported that principal investigators included data management in their grant proposal budget. Still, this number was higher than the 13.95 percent who said they did so in the initial survey. The majority of respondents (31 percent) indicated a hybrid model of funding, drawing from a mixture of all sources. Also important to note is that twice as much funding apparently comes from the library budget than from any other department budget.

Discussions in focus groups reflected this state of affairs; participants reported that vice presidents of research and other administration officials are reluctant to commit funding or other institutional resources to research data management support. Participants described conversations with administrators who believed the NSF mandate was just a phase, who expected those in the disciplines to revolt and simply stop reviewing plans until the mandate went away, and who would not invest in research data management support because the "return on investment" was unclear. The latter position is fundamentally anti-intellectual and reveals a deep misunderstanding of the basic principles of research, which ideally begins with an unanswered question, not a financial statement. This is, unfortunately, perfectly consistent with the technocratic logic driving many university administrations.

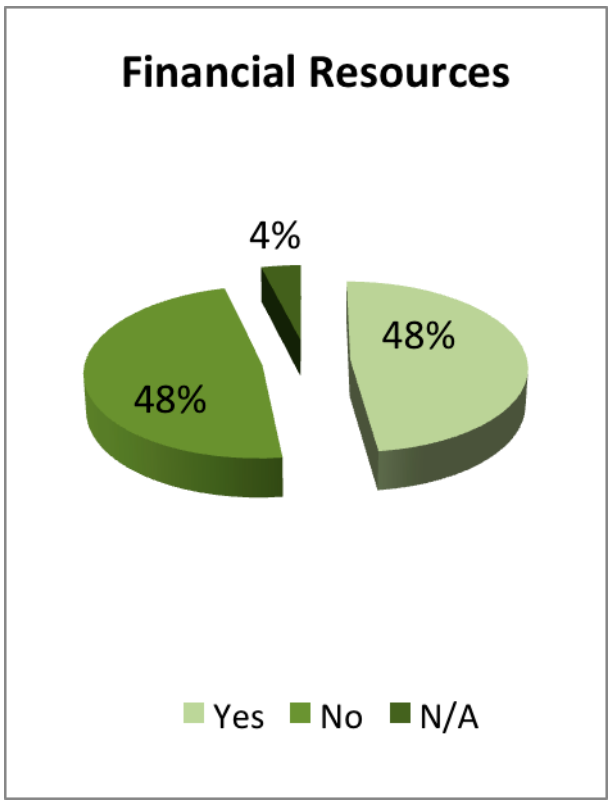

Fig. 9. Administrator responses to survey question, "Does your institution allocate financial resources for data management?"

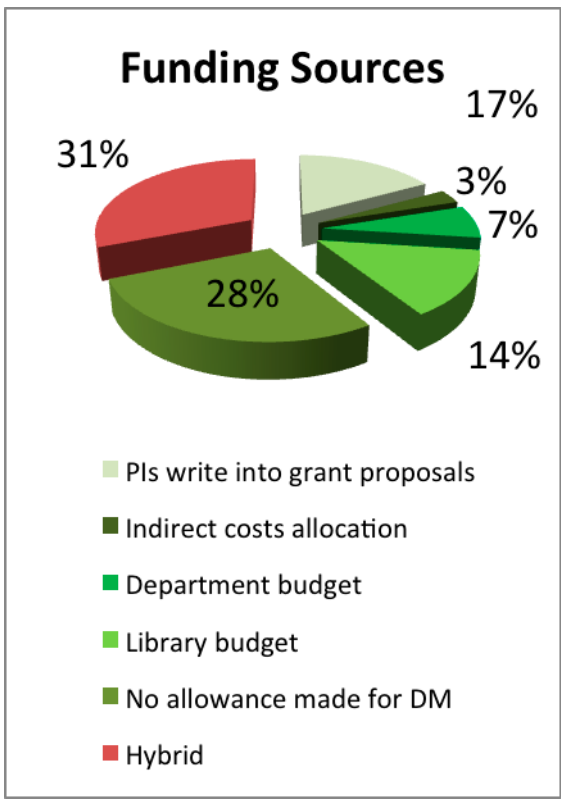

Fig. 10. Administrator responses to survey question, "From which sources are these funds drawn?" Multiple answer choices could be selected.

Conclusion: The Data Doldrums

At our final focus groups, conducted in June 2013 during the American Library Association annual conference in Chicago, we asked participants (mostly academic librarians) what message they most wanted us to take away from the discussion. Early in the research process, focus group participants had been anxious about the cost of implementing data management services, but eager to hear what was happening at other institutions and to share gossip and anecdotes about badly behaved principal investigators and administrators. In stark contrast, the atmosphere in the Chicago focus groups was noticeably subdued. Participants in these later focus groups most often used words like "worried," "anxious," and "stressed" to describe their feelings about data management services at their institutions.

One participant, a liaison librarian from a prestigious private research university, eloquently expressed her fear that library support for disciplines like philosophy, the humanities, and the soft social sciences would be left behind as university administrations and offices of research, library leadership, and funding agencies, including NEH–ODH, turn away from supporting traditional lines of scholarly inquiry in favor of data-driven (in particular, big data-driven) projects that are now "sexy." Without new funding to support research data management functions, the new focus on research data management will likely end up further overloading already overstressed

library budgets. This could potentially threaten to further weaken support for non-STEM (science, technology, engineering, and mathematics) research in favor of funding agency and university administration priorities that in many cases researchers in the STEM disciplines do not (yet) share.

How then do we finally understand the current status of research data management efforts in academia? In the two and a half years since NSF announced its data management plan requirement, academic libraries have scrambled to keep up with what continues to be perceived as another unfunded mandate. Returning to the nautical metaphors popular in discussions of big data, we are neither riding the wave nor being swamped by it. Rather, we may be *becalmed*, mired in the Sargassum of institutional inertia.

There was a significant degree of hope that the February 2013 memo of the Office of Science and Technology Policy nudging federal agencies to come up with a coherent strategy would spark some movement, but the August deadline for agency plans came and went with no public announcements. This silence was soon followed by the shutdown of the U.S. federal government in October 2013, an event that is all too emblematic of gridlock and being stuck in the doldrums. It now seems highly unlikely that vigorous and assertive prescriptions for research data management will be forthcoming from federal agencies, at least in the immediately foreseeable future.

In the absence of clear guidance from the federal agencies, university administrations are likely to fall back on the all too easy excuses for withholding resources from service providers—mainly libraries—believing that the requirement for a data management plan is a passing whim on the part of the agencies and that there is no point in investing time, money, and staff without a clear return on investment. Principal investigators, too, have room to doubt the seriousness of the data-sharing mandate and may simply continue to craft data management plans that reinforce the proprietary nature of their data rather than planning to make them available to be preserved, shared, and repurposed. And libraries may continue to try to meet the demands of both administrators and researchers with ever-shrinking financial resources—the equivalent of diligently polishing the decks and patching the sails in the vain hope that today the winds will return.

We should not allow our institutions to continue to drift in the data doldrums. To continue our metaphor, the promise of new lands is too great for us to accept remaining becalmed. But if we are to emerge from the doldrums, we will have to demonstrate stronger leadership and make greater efforts to work together within our institutions. Rather than waiting passively, we should take serious analytic notes from the small number of exemplar institutions in which librarians, researchers, and academic administrative leaders are collaboratively developing a shared agenda for research data management. Rowing out of the doldrums will require hard work, and we will have to row *together* to succeed. The question is not really

whether we will devote this effort to moving forward, it is rather *how long* we will collectively tolerate being becalmed. Our conclusion is that we should collectively *get moving*.

References

California Digital Library. 2013a. DMPTool. Available at https://dmp.cdlib.org/.

California Digital Library. 2013b. EZID. Available at http://www.cdlib.org/services/uc3/ezid/index.html.

California Digital Library. 2013c. University of California Curation Center. Available at http://www.cdlib.org/services/uc3/index.html.

Hermeneutica.ca. n.d. Voyant. Available at http://voyant-tools.org/.

Institute for Museum and Library Services. n.d. Specifications for Projects That Develop Digital Products. Available at http://www.imls.gov/assets/1/AssetManager/DigitalProducts.pdf.

Kelley, Michael. 2011. Librarians at University of Minnesota Make an Impact with Data Management Program. *Library Journal* (Aug. 9).

MIT Libraries. n.d.a. DSpace@MIT Available at http://dspace.mit.edu/.

MIT Libraries. n.d.b. *Self Help: Subject Guides: Data Management and Publishing*. Available at http://libraries.mit.edu/guides/subjects/data-management/.

Moen, William E., and Martin Halbert. 2012. *Support for Research Data Management among U.S. Academic Institutions: Results from a National Survey*. Denton: Texas Center for Digital Knowledge, College of Information, University of North Texas. Unpublished draft dated January 26, 2012; cited by permission. UNT Digital Library. http://digital.library.unt.edu/ark:/67531/metadc181695/.

Moretti, Franco. "Literary Lab Pamphlet 2: Network Theory, Plot Analysis." Stanford: Literary Lab, May 1, 2011. Available at http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf.

National Endowment for the Humanities. Data Management Plans for NEH Office of Digital Humanities Proposals and Awards. Executive Summary. Available at http://www.neh.gov/files/grants/data_management_plans_2013.pdf

National Institutes of Health. 2003, February 26. Final NIH Statement on Sharing Research Data. NOT-OD-03-032. Available at http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

National Science Foundation. n.d. Dissemination and Sharing of Research Data. Available at http://www.nsf.gov/bfa/dias/policy/dmp.jsp.

National Science Foundation. 2011. *Award and Administration Guide*. Chapter VI.D.4. Available at http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4.

National Science Foundation. 2013. *Grant Proposal Guide*, Chapter II.C.2.j. Available at http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp.

Pennsylvania State University Libraries. Digital Curation Services. Available at http://www.libraries.psu.edu/psul/pubcur/curation.html.

Purdue University Libraries. n.d. Data Curation Profiles Toolkit. Available at http://datacurationprofiles.org/.

Purdue University. 2013. Purdue University Research Repository. Available at https://purr.purdue.edu/.

Tufts University, 2013. Research Guides@Tufts. *Federal Funding Agencies: Data Management and Sharing Policies*. Available at http://researchguides.library.tufts.edu/content.php?pid=167647&sid=1412586.

University of California at San Diego. 2013a. RCI: Research Cyberinfrastructure. Available at http://rci.ucsd.edu/.

University of California at San Diego, The Library. 2013b. Research Data Curation Services. Available at http://libraries.ucsd.edu/services/data-curation/index.html.

University of Illinois at Urbana-Champaign. n.d. Illinois Digital Environment for Access to Learning and Scholarship. Available at https://www.ideals.illinois.edu/.

University of Illinois at Urbana-Champaign, University Library. Life Sciences Data Services, Data Management Plans. Available at http://www.library.illinois.edu/lsdata/dmp/dmp.html.

University of Minnesota, University Libraries. 2011. Managing Your Data. Available at http://www.lib.umn.edu/datamanagement.

University of New Hampshire. (2012) Ownership and Management of Research Data. UNH.VII.C.1 http://www.usnh.edu/olpm/UNH/VIII.Res/C.htm.

The Denton Declaration: An Open Data Manifesto

Introduction

On May 22, 2012, at the University of North Texas, a group of technologists and librarians, scholars and researchers, university administrators, and other stakeholders gathered to discuss and articulate best practices and emerging trends in research data management. This declaration bridges the converging interests of these stakeholders and promotes collaboration, transparency, and accountability across organizational and disciplinary boundaries.

Declarations

- Open access to research data is critical for advancing science, scholarship, and society.
- Research data, when repurposed, has an accretive value.
- Publicly funded research should be publicly available for public good.
- Transparency in research is essential to sustain the public trust.
- The validation of research data by the peer community is an essential function of the responsible conduct of research.
- Managing research data is the responsibility of a broad community of stakeholders including researchers, funders, institutions, libraries, archivists, and the public.

Principles

- Open access to research data benefits society, and facilitates decision making for public policy.
- Publicly available research data helps promote a more cost-effective and efficient research environment by reducing redundancy of efforts.

- Access to research data ensures transparency in the deployment of public funds for research and helps safeguard public goodwill toward research.
- Open access to research data facilitates validation of research results, allows data to be improved by identifying errors, and enables the reuse and analysis of legacy data using new techniques developed through advances and changing perceptions.
- Funding entities should support reliable long-term access to research data as a component of research grants due to the benefits that accrue from the availability of research data.
- Data preservation should involve sufficient identifying characteristics and descriptive information so that others besides the data producer can use and analyze the data.
- Data should be made available in a timely manner; neither too soon to ensure that researchers benefit from their labor, nor too late to allow for verification of the results.
- A reasonable plan for the disposition of research data should be established as part of data management planning, rather than arbitrarily claiming the need for preservation in perpetuity.
- Open access to research data should be a central goal of the lifecycle approach to data management, with consideration given at each stage of the data lifecycle to what metadata, data architecture, and infrastructure will be necessary to support data discoverability, accessibility, and long-term stewardship.
- The costs of cyberinfrastructure should be distributed among the stakeholders—including researchers, agencies, and institutions—in a way that supports a long-term strategy for research data acquisition, collection, preservation, and access.
- The academy should adapt existing frameworks for tenure and promotion, and merit-based incentives to account for alternative forms of publication and research output including data papers, public data sets, and digital products. Value inheres in data as a standalone research output.
- The principles of open access should not be in conflict with the intellectual property rights of researchers, and a culture of citation and acknowledgment should be cultivated rigorously and conscientiously among all practitioners.
- Open access should not compromise the confidentiality of research subjects, and will comply with principles of data security defined by HIPAA, FERPA, and other privacy guidelines.

Intentions

In our professional interactions at meetings, on review panels, conferences, teaching, etc. we will advocate the following positions:

- A culture of openness in research.
- A federated model of archiving data to enable discoverability, transparency, and open access.
- A robust and sustainable funding regime for research data management infrastructure (technical, policy, and human resources).

- The development and adoption of metadata standards for research data.
- Long-term access to data that supports published research outputs.
- Support for researchers in negotiations with publishers to allow open access to research in repositories.
- Recognition of researchers' intellectual property in data and scholarly research outputs.

Invitation

We invite all others who support these principles of research data management to join with us to make our vision of a culture of open data a reality.

Join us! Add your support to the principles of open data by adding your signature at http://openaccess.unt.edu/denton-declaration. Organizations wishing to lend their support, please email datares@unt.edu.

Participants

Jonathan Crabtree, Assistant Director for Archives and Information Technology, H.W. Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill

Stephen Griffin, Professor in Cyberscholarship, School of Information Sciences, University of Pittsburgh

Michael Greenlee, Reference, Instruction, and Web Services Librarian, University of Tulsa

José-Marie Griffiths, Vice President for Academic Affairs, Bryant University; National Science Board

Martin Halbert, Dean of Libraries, University of North Texas

Michael Hulsey, Technical Applications Specialist, Immunocytometry Systems Group, BD Biosciences

James H. Kennedy, Regents Professor and Director, Elm Fork Education Center and Natural Heritage Museum, Department of Biological Sciences, University of North Texas

Spencer D. C. Keralis, Director for Digital Scholarship, University of North Texas

John Kunze, Associate Director, University of California Curation Center

William E. Moen, Associate Dean for Research, College of Information, University of North Texas

Allen Renear, Professor, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Kenneth W. Sewell, Associate Vice President for Research, University of North Texas

Brian E. C. Schottlaender, The Audrey Geisel University Librarian, University of California – San Diego

Denise Perry-Simmons, Assistant Vice President, Research Development, University of North Texas

Shannon Stark, Strategic Projects Librarian, University of North Texas

Carly Strasser, Project Manager, Data Curation for Excel, California Digital Library

Rene Tanner, Life Sciences Librarian, Arizona State University

Why, How, and Where We're Going Next: A Multi-Institution Look at Data Management Service

Kiyomi Deards

Abstract

This study compares the relationship of four university libraries with their institutions' data management policies, services, resources, and plans for the future. It combines an examination of public documentation and services advertised by the universities and their libraries with interviews of librarians involved with data management. Four members of the Association of Public and Land-Grant Universities of similar size were selected for comparison: Iowa State University, Oklahoma State University, Kansas State University, and the University of Nebraska–Lincoln. The study identifies how and why campus collaborations were created and describes possible next steps for expanding or refining existing data management services. University-wide data management resources and services vary significantly depending on the size of the institution and other factors, such as the commitment of campus administration and faculty buy-in. This study suggests best practices and methods of collaboration between libraries and the universities that they serve.

Introduction

The data curation and management services planned for and offered by four universities selected from the Association of Public and Land-Grant Universities (APLU) are examined in this research. The term *data curation and management services* is defined here as any services related to the organization, management, or long-term preservation of data developed through scholarly research. These services encompass a range of activities, including consultations on creating data management plans and strategies, physical or electronic archiving of datasets, and workshops.

Research Design

This section describes the selection of participants for this study, how their websites were analyzed, and who was selected to be interviewed. It also establishes the timeline of this research to enable readers to determine how websites, services, and personal have changed since the writing of this article.

Selection of Participants

The U.S. Department of Education listed 9,675 unique post-secondary institutions in the United States in its December 2012 database. Of these, only 105 are land-grants, 74 of which are APLU members representing approximately 7.6 percent of accredited post-secondary institutions (Association of Public and Land-Grant Universities 2012; Smithsonian Institution 2012). Land-grant institutions differ from other state schools in that their mission is to teach agricultural, martial, mechanical, and classical studies (Washington State University Extension 2009). They are home to extension services, which provide research and advice to advance a state's agricultural endeavors. Consequently, extension offices generate large amounts of data and publications, creating a need for data curation and management.

Four land-grant universities were selected for comparison of their data curation and management services: Iowa State University (student population of 29,887), Kansas State University (student population of 23,863), Oklahoma State University (student population of 21,419), and the University of Nebraska–Lincoln (student population of 24,593). They were selected primarily because of their proximity, similarity of mission as land-grant members of APLU, and their relative similarities in student body size. It was theorized that similarly sized universities with similar missions would develop services related to data curation and management that would be transferable between institutions.

Website Comparison Process

Universities' public websites were found, using Google, and examined to see what services the libraries and their universities offered with respect to data curation and management. Web searches were conducted, using Google, on the libraries' websites and on each university's website. In addition, university library and office of research websites were browsed manually to explore where services were listed and how easy they were to locate. Types of services, number of related pages, number of contact people listed, and observations on ease of location of data services were recorded.

Interview Contact Process

An e-mail sent to the contact people listed on each university's website asked recipients to complete a phone or e-mail interview. Three of the four institutions listed a contact person associated with the library who either chose to participate or referred the researcher to

a more appropriate person who chose to participate. One participant referred the researcher to the original organizer of the library's data curation and management efforts, who had changed universities, and that person also agreed to participate. Iowa State listed no library services related to data curation and management, and no contact person; no one was interviewed from this university. All of the individuals who participated asked to be interviewed by e-mail because of time constraints. The purpose of these interviews was to gain perspective on how and why data services were developed and what services have been planned for the future. The questions were kept short and designed to elicit information that the researcher did not expect to find on the library or university websites. All interviewees were asked to address the same questions (see Appendix, E-mail Interview Script).

Timeframe

Interview planning took place in August 2012 with institutional review board (IRB) approval granted through the University of Nebraska–Lincoln (UNL) on September 7, 2012 IRB #: 20120912867 EX. Website review began September 7, 2012, and the initial e-mail requests for interviews went out the same day. Website analysis continued through October 2, 2012. Responses to the e-mail surveys were received by October 2, 2012, and the analysis was concluded by October 31, 2012. Results of this research were presented at the December 10, 2012, DataRes Symposium of the Coalition for Networked Information (CNI) in Washington, D.C.

Website Comparison of Services

Initially, Google was used to search for the key word *data,* along with the university's name using Google and the university websites. These searches yielded thousands of results for information services technical departments and computer science publications and classes, burying the results on data curation and management services. Each university website was then searched using the internal search function to identify the office of research or equivalent website, and university library's website. The search also identified web pages and resources related to data curation and management services within the university website. The researcher used the Google search tool to do two searches: one for "data curation" and one for "data management"; both incorporated the name of the university. This change in search strategy eliminated many of the irrelevant search results. Fifty pages of search results were examined for each university and search phrase. The researcher posited that those looking for data curation and management guidance would not view more than 50 pages of results. Lastly, all university websites were searched using the term *data libraries* in order to identify additional connections between the university libraries and data curation and management services.

The office of research or equivalent body at all four universities

provided information related to data management plans (table 1). Iowa State, Oklahoma State, and UNL had departments or centers of research that provided information on data curation or data management on their website. Two universities, Kansas State and Oklahoma State, listed a librarian contact for data curation or management services. At Oklahoma State, this librarian was listed on a non-library web page of the university. UNL provided services via committee. No library contact or services were listed on the Iowa State library website.

Kansas State and UNL provided information on their main website; they offered a library-maintained guide to data curation and management, and related consultants (table 1). Iowa State and Oklahoma State did not list services or librarians related to data management on the library portion of their websites. It was found that Oklahoma State provides data archiving services through an associated state university for a fee, while UNL provides data archiving services to faculty, staff, and students for a fee. Iowa State and Kansas State University did not list archiving services for faculty, staff, and students on their websites. The size of the student body did not correlate to the number or type of services offered.

Websites, Services Offered, & Positions Related to Data	Iowa State University	Kansas State University	Oklahoma State University	University of Nebraska–Lincoln
Number of students	29,887	23,863	21,419	24,593
Office of research or equivalent	X	X	X	X
Department(s)/center(s) of research	X		X	X
Library information on main site		X		X
Library-maintained guide		X		X
Data curation / management consultants		X		X
University-provided data archiving services				X
Data archiving services through associated university			X	
Library committee				X
Librarian contact		X	X	

Table 1. Websites, services offered, and positions related to data service

Data curation and management advice and services were listed on a total of 47 web pages and tabs on the university and library websites (table 2). Tabs were counted along with web pages when clicking the tab changed the primary information on the web page. Kansas State and UNL had the most related web pages and tabs with 13 and 25, respectively (table 2). This correlated to the higher number of services offered by these universities, as shown in table 1. Oklahoma State University had six, and Iowa State University had three, correlating to the lower number of related services listed by those two universities and their libraries.

UNL had information on 10 web pages and tabs controlled by the Libraries and on 15 throughout the website. Kansas State had information on seven library web pages and tabs, and an additional six throughout the website. Oklahoma State had no information listed on its library web page, but services by the libraries were listed on the related university web pages (tables 1 and 2). No related library services were listed on the Iowa State website (table 1).

	Iowa State University	Kansas State University	Oklahoma State University	University of Nebraska-Lincoln
Library	0	7	0	10
Rest of the university	3	6	6	15
Total	3	13	6	25

Table 2. Number of pages or tabs on website referring to data curation and management

The number of clicks and hovers needed to reach relevant information determines how difficult it is for users to reach the information that they need. Hovering over drop-down and side menus, and then clicking on a link was counted as two steps. Multi-tabbed web pages are treated as one primary web page in table 3. Files found on these pages are not counted in these statistics. Pages found through a Google or website search were counted as one step if only one click was needed to reach the relevant information.

Access to primary data curation and management web pages generally required one to two clicks: 17 and 13 times, respectively (table 3). Only once were three clicks required, and only once were four clicks necessary. This demonstrates that information on data curation and management is being assigned priority as top and secondary level content.

Number of Steps	Iowa State University	Kansas State University	Oklahoma State University	University of Nebraska-Lincoln	Total
1	0	4	1	12	17
2	3	1	4	5	13
3	0	0	0	1	1
4	0	1	0	0	1

Table 3. Number of steps to reach a relevant page on the websites

The Kansas State and UNL web pages linked to multiple outside resources on data curation and management (table 4). These resources included the Inter-university Consortium for Political and Social Research (ICPSR) website, the National Science Foundation (NSF) website, the National Institutes of Health (NIH) website, templates and sample text, and data repositories. Iowa State's web pages linked to the NSF website. Oklahoma State's website linked to the NIH and NSF websites. All the universities linked to the NSF website,

demonstrating the emphasis on NSF data management plans and guidelines. Three of the universities (75%) linked to the NIH website; two (50%) to the ICPSR website. This trend corresponds to the amount of money given in grants by these groups; funding from the NSF is the highest, followed by that from the NIH, and trailed by the funding available from agencies for the social sciences.

	Iowa State University	Kansas State University	Oklahoma State University	University of Nebraska–Lincoln
ICPSR website		X		X
NIH website		X	X	X
NSF website	X	X	X	X
Templates / sample text		X		X
Data repositories		X		X

Table 4. University links to outside resources

Oklahoma State and UNL both provide data storage solutions for faculty. Oklahoma State provided storage through the PetaStore hosted at the University of Oklahoma, with services available to all institutions of higher education in the state (Neeman and Calhoun 2012). UNL hosted its own Data Store and offered services only to individuals with UNL identification (ID) cards (University of Nebraska–Lincoln 2012m). One terabyte (TB) is equal to 1,000 gigabytes. Assuming the cost per terabyte holds steady, the cost to store, back up, and preserve 1.5 TB of information is $300 at Oklahoma State and $7,500 at UNL (table 5). Neither data storage solution guarantees that file formats will be updated as time goes by; they ensure only that the data will remain preserved. Additionally, no mention was made of a secure archive for potentially sensitive information that must be not only preserved, but also guarded against accidental disclosure. Although the use of 1.5 TB tapes may impede someone looking to steal specific research information, no procedures are currently in place to safeguard and preserve information generated by individuals studying at-risk populations. No costs for data storage were listed on the Iowa State or Kansas State websites.

	Iowa State University	Kansas State University	Oklahoma State University	University of Nebraska–Lincoln
Cost of services / amount of data	None listed	None listed	$150 / 1.5-TB tape cartridge (2 tapes are recommended for backup and preservation)	$500/100 GB $1,250/200 GB $2,500/500 GB $5,000/1,000 GB

Table 5. Cost of Data Management Services Offered by Institution or Consortia

Details of Website Discovery

Unless otherwise mentioned, searchers needed to click only once on the relevant web page to access information on data curation or management on either the university website or the results page of a Google search. The need to click more than one link to access relevant information increased the likelihood that searchers would become frustrated or give up searching. Links to NSF, NIH, and other funding agencies demonstrated an awareness of who is requiring data curation and management, and the possibility that the requirements established by these agencies can be updated at any time.

Iowa State University

Iowa State had two web pages with information related to data management plans. First, the Office of Sponsored Programs Administration homepage contained a link to NSF Data Management Plan Info (Iowa State University 2012). Selecting "Sponsor Requirements" and then "NSF Data Management Plan Info" led to "National Science Foundation Data Management Plan Requirements" and "Creating Data Management Plans for NSF Proposals: Template" (Iowa State University 2012; Iowa State University a, b).

Second, the College of Agriculture and Life Sciences devoted space to data management plans under "CALS Funding Resources: Tips and Tools" (College of Agriculture and Life Sciences 2012). That led to a webpage with the same information contained in the "CALS Funding Resources: Tips and Tools" and a downloadable document, "Developing NSF Data Management Plans" (College of Agriculture and Life Sciences 2012). When opened, the file was titled "Suggested Practices: Developing a Data Management Plan for NSF Grant Applications" (Clemens 2011). Links to the NSF frequently asked questions (FAQs), data management plan development resources and examples, and information on what costs to include in data management plans were provided in the document.

No other relevant web pages, or services, were discovered on the Iowa State website. Searching the website for "data libraries" resulted in no additional relevant search results, nor were any additional pages discovered by searching for more information using Google.

Kansas State University

Kansas State had multiple websites and services dealing with data curation and management services. The Office of Research and Sponsored Programs listed two resources on its "Proposal Writing Resources" page: "NIH Data Management Plans" and "NSF Data Management Plans" (Kansas State University Office of Sponsored Programs 2012f). "NIH Data Management Plans" provided information on NIH's data sharing requirements and linked to NIH resources (Kansas State University Office of Sponsored Programs 2012c). "NSF Data Management Plans" provided examples of information required and speculated about possible future requirements (Kansas State University Office of Sponsored Programs 2012d). It linked

to the "K-State Libraries Data Management Planning" website, the Libraries' contact person, and multiple planning tools and example text (Kansas State University Office of Sponsored Programs 2012d). The guidance link suggested strategies and justification for data management, and it listed available library and university services (Kansas State University Office of Sponsored Programs 2012e).

The Workshops and Training icon led to "Classes, Workshops, and Training by Date" (Kansas State University Office of Research and Sponsored Programs 2012a). On May 3, 2012 a Data Management Workshop was held (Kansas State University Office of Research and Sponsored Programs 2012b).

Searching the Kansas State website for information on data management led to the Kansas State Data Commons web page. It provided information on data sources; data archives; and Kansas State, state, and regional resources related to data (Kansas State University 2012b).

From the "Libraries" main page, it took four clicks—(1) "About Us," (2) "Departments," (3) "Divisions and Departments," and (4) "Faculty and Graduate Services—to discover "Data Resources and Services" (Kansas State University 2012a). Contact information for Data Resources and Services, and a link to the "Data Management Research Guide" appeared there (Kansas State University 2012a). The guide explained why researchers should manage their data (Duever 2012a). It also gave contact information for the Data Services Librarian, provided a PDF guide to data management, and linked to the ICPSR website. The "Organizing Your Data" tab focused on the organization and sustainability of data, and it linked to resources about organizing data (Duever 2012e). The "Metadata" page defined the term *metadata* and provided standards and instructions on how to use it (Duever 2012d). The "Archiving/Preservation" page contained information on why and how to back up data, and linked to data repositories (Duever 2012b). "Sharing Your Data" explained why, where, and how researchers should share data (Duever 2012f). "Data Management Resources" included links to checklists, planning tools, Kansas State resources related to data management, and data repositories (Duever 2012c).

Searching the Kansas State website for "data libraries" produced no additional relevant search results. No additional websites were discovered by searching for more information using Google.

Oklahoma State University

The University Center for Proposal Development at Oklahoma State University had a link to "NSF Data Management Plans" from its "Resources" web page (Oklahoma State University 2012b). Two links to local resources did not work. "NSF DMP [Data Management Plan] Requirements and Guidance Document" led to a PDF about NSF data management plans, data repositories, the Oklahoma PetaStore, and NIH's "Data Sharing Policy and Information Resources" website (Oklahoma State University Office of Proposal Development 2011). Hovering over "Resources—NSF Data Management Plans" revealed

the "DMP Suggested Structure" web page. The page broke data management plans into six sections with suggested practices for each section (Oklahoma State University b).

"Newsletter, Special Events" advertised a 2011 NSF Regional Grant Conference that covered data management plans (Oklahoma State University 2011). "Newsletter, In the Know" had sample text that a researcher could copy into a data management plan with modification (Oklahoma State University 2012a). The text referred to the Oklahoma PetaStore, which provides tape storage at $150 per 1.5-TB tape cartridge (Oklahoma State University 2012d). Two tapes were to be purchased, one for storage and one as a backup.

Oklahoma State's website was searched for "PetaStore," leading to the discovery of "NSF Data Management and Sharing Plans." The document explained the necessity of data management plans and their requirements, and it linked to online resources (Oklahoma State University 2012c). It also described the Oklahoma State University Library as having limited opportunities for data archiving. The University Research Services "Helpful Links" web page also linked to this document (Oklahoma State University a).

A Google search for "PetaStore Oklahoma State University" revealed that the PetaStore at the University of Oklahoma had been opened for use in October 2011 (Buckmaster 2011). Search results included the "University of Oklahoma Supercomputing Center for Education and Research" homepage, which contained information about the PetaStore (University of Oklahoma 2012a). "Accounts—Oklahoma State PetaStore Policies and Procedures" revealed information on the PetaStore functions and a link to the "PetaStore Use Agreement" (University of Oklahoma 2012b). Lastly, a 2012 presentation for the Great Plains Network Annual Meeting was found. The presentation indicated that the PetaStore is open to all types of institutions, but it is free only for Oklahoma academic users (Neeman and Calhoun 2012).

Searching Oklahoma State's website for "data libraries" resulted in no additional relevant search results.

University of Nebraska–Lincoln

The website of the Office of Research and Economic Development of UNL had three links leading to two pages with data management plan information (University of Nebraska–Lincoln 2012f). Selecting the "Office of Research—Proposal Development" or "Our Offices—Proposal Development" led to the same website (University of Nebraska–Lincoln 2012f). "Proposal Development" contained a link, under "Resources," to "NSF Data Management Plan Resources" (University of Nebraska–Lincoln 2012h). "For Researchers—Faculty Resources" led to "NSF Data Management Plan Resources" (University of Nebraska–Lincoln 2012f, 2012g).

"NSF Data Management Plan Resources" provided an overview of NSF requirements and linked to "NSF Division and Directorate Policies" (University of Nebraska–Lincoln 2012i). An outline that was provided online could be downloaded, and review services were

offered. The Data Curation Working Group web page was listed as an additional resource, as were other universities' online resources (University of Nebraska–Lincoln 2012i).

"What's New" was accessed via the main library website drop-down menu (University of Nebraska–Lincoln 2012n). A list of Spring Semester 2013 lectures included a lecture with the title "Data Management Basics."

"About" on the UNL Libraries website led to "Data Management Plans" (University of Nebraska–Lincoln 2012k). The "Data Management" page had five tabs:

1. "Write a Plan" gave an overview, linked to specifics of various agency requirements, and provided a downloadable template and a link to request help (University of Nebraska–Lincoln 2012d).
2. "Examples" provided sample data management plans, sample text, and a link to request help (University of Nebraska–Lincoln 2012d).
3. "Checklist" had a list of questions for researchers to address when preparing a data management plan and a link to request help (University of Nebraska–Lincoln 2012d).
4. "Workshops & Consultations" contained information on services provided and a link to request help (University of Nebraska–Lincoln 2012d).
5. "Deposit" led to the UNL Data Repository website (University of Nebraska–Lincoln 2012d, 2012m).

The "UNL Data Repository" homepage featured information about data storage services and the related costs, starting at $500 for the storage of 100 GB (University of Nebraska–Lincoln 2012m). Information on repository contacts was also supplied. There was a link to begin depositing data; a university ID and e-mail address were required to register and view the interface (University of Nebraska–Lincoln 2012m; UNL Data Repository 2011).

The "UNL University Libraries Guide to Data Management" was located from the main library website by selecting "Subject & Course Guides," "Data Management," and "Data Management" again. The main page explained data management, linked to the library data management website, gave contact information for the guide maintainer, and linked to a book on data management (Deards 2012a).

"Management & Preservation" provided information on writing a data management plan, sample plans and templates, boilerplate text, storage, sustainability, and links to information on data management plans from NSF directorates (Deards 2012e). The sidebar contained a checklist, information on arranging consultations and workshops, a fun data fact, and links to other institutions' data management plan information pages.

"Documentation & Storage" defined the term *metadata*, examined file-naming conventions, and explained and linked to controlled vocabularies (Deards 2012c). Data storage and backup, using the

UNL Data Store, security issues, and sustainable data formats were also addressed.

"Need Help" contained links to e-mail for help and to the Office of Research and Economic Development's proposal development page on "NSF Data Management Plan Resources" (Deards 2012f). It also linked to other academic and professional guides to creating and managing data management plans.

The section on FAQs addressed the impact of government sponsorship, summarized the role of libraries and archives, and provided contact information for those working in data curation (Deards 2012d). It described the Libraries' partnership with Information Services to create the Data Store, and it defined the Data Curation Working Group's purpose. "Definitions" presented definitions and links to cited references for nine data management terms (Deards 2012b).

The Google searches revealed four presentations and news announcements. One presentation covered the importance of data curation and the services developed to address researchers' needs by the Data Curation Working Group (Westbrooks 2011). A presentation abstract, titled Archival Data Management for Research**,** was linked to a Prezi presentation (Coalition for Networked Information 2011; Westbrooks and Notter 2011). The presentation focused on the elements needed to develop the UNL Data Repository and the next steps in its development. "The Scarlet," a news source, highlighted the libraries' help with data management plan development (tfedderson2 2011). "Research News" highlighted the NSF data management plan requirements and the related services offered by the Office of Proposal Development and the Libraries (UNL Office of Research 2011).

Searching the UNL website for "data libraries" led to one conference poster presentation about the Data Store, two audio recordings, a handout, and two news posts. "Preserving the Present to Inform the Future: Issues in Data Preservation and Access" was the conference poster presentation (Deards 2012g). "Data Curation—Academic Activities Brown Bag" was a discussion led by the Data Curation Working Group (University of Nebraska–Lincoln 2012a). The "Data Curation FAQ Session and Discussion" was led by Elaine Westbrooks (University of Nebraska–Lincoln 2012b). "The Experienced Viewpoint: Advice on Digital Data Storage and Management Suggestions and Comments from UNL Library faculty Scott Childers and Leslie Delserone" was posted on UNL's Water Center website (Childers and Delserone 2011). "Today@UNL," the campus news website, highlighted the libraries' data management plan website, as well as the consultation and workshop services for faculty on September 15, 2011 (University of Nebraska–Lincoln 2012e). Also found were "IANR Land Grant Legacy Celebration," which highlighted data management services and partnerships, and "University Libraries Strategic Plan 2012–2013" (University of Nebraska–Lincoln 2012c, 2012j). Assisting faculty with data management plan development was part of "Priority 3. Advance the Libraries' role in supporting

research programs/scholarly and creative activities" (University of Nebraska–Lincoln 2012j).

Results of Interviews with Data Curation and Management Services Library Leaders

Four individuals, representing three institutions, responded to six open-ended questions regarding data curation and management services at their current or former university. All respondents identified data curation and management as important areas that would only become more important over time. They identified workshops for faculty and individual consultations regarding data management plans as successful services. Librarians alone or in collaboration with representatives from the office of research and information services led the development of data-related services. Even when librarians alone led the development of services, personnel from the office of research and information services were sought out as vital campus collaborators. Respondents viewed data management as critical to the continued advancement of science, the digital humanities, social sciences, and other areas of research that generate large amounts of data. They listed the lack of resources, funding, staff, and technology as the greatest barriers to the creation and success of services. Future plans include increasing instruction in data management plans to graduate students, refining data preservation services, and consulting with stakeholders to tailor services to their needs and create buy-in.

The following provides more specific information gathered from the interviews with data curation and management services library leaders. Numbers are shown in the format (1/4), (2/4), (3/4), and (4/4) to indicate how many respondents made similar statements.

Campus Collaborators

Three of the four respondents identified the Office of Research, or its equivalent, and the Computing (or Information Service) Department as the most important campus collaborators in developing in-house data management services. The fourth respondent did not identify the Office of Research as a potential partner, but did identify it as a source of information on the campus need for data management services.

- Office of Research or equivalent (3/4)
- Computing or Information Services Department (3/4)

Services Offered

In three of the four universities, libraries had initially developed the existing services; in the other university, the library was solely responsible for the development. All respondents noted that their libraries provided consultations on data management plans. Three of the four provided evaluations of data management plans and instruction in data management best practices. Two of the libraries provided data storage through an institutional repository. These

results show a focus on creation of data management plans and less emphasis on an institutional repository.

- Data management plan consultations (4/4)
- Evaluations of data management plans (3/4)
- Instruction in best practices in data management (3/4)
- Data storage in an institutional repository (2/4)

Perceived Benefits of Data Services

All respondents viewed data services as meeting the needs of researchers, supporting the mission of the libraries, and increasing the value of the libraries to the university. Two of the respondents thought that services help faculty think about future reuse of their data, help bring in more funding, and advance the prestige of the university. One of four respondents posited that data services may help create better datasets for the future.

- Meets the needs of researchers created by the NSF data management plan requirements (4/4)
- Supports the libraries' mission to support researchers and their work (4/4)
- Increases the perceived value of the library to the university (4/4)
- Helps faculty to explore ways that their data may be used in the future and consider their options for data preservation (2/4)
- Supports researchers in their pursuit of funding, which brings in more money to the campus (2/4)
- Advances the prestige of the university as a leader in research (2/4)
- Creates better datasets for future use (1/4)

Perceived Barriers to Data Services

Money, time, and lack of buy-in were cited by three of the respondents as barriers to creating and maintaining successful data curation and management services. Two respondents noted a lack of successful communication between collaborators. Mentioned once by respondents as problems were inadequate staff, lack of clarity about which campus units should collaborate, and difficulty of sustaining services. One respondent noted the perception of libraries as a place to contact with problems, not for expertise. One respondent also posited that faculty may choose to use established disciplinary repositories over newer and less proven institutional repositories.

- Money (3/4)
- Time needed to develop and maintain data-related services (3/4)
- Lack of buy-in (3/4)
- Lack of successful communication between collaborators (2/4)
- Lack of staff devoted to full-time data services (1/4)
- Lack of understanding about how campus units should work together to support successful data management and preservation by researchers (1/4)
- Lack of sustainability of current data archives (1/4)
- Perception of libraries as a place to turn only if you are having a problem, not as a place of expertise to consult on a regular basis (1/4)
- Faculty choice to use older more established subject repositories over institutional repositories (1/4)

Proposed Future Steps

Respondents' planned future steps were varied. Two of the four intended to speak with senior administrators to build support for the development of services. One respondent proposed workshops for graduate students, and another suggested refining policies for internal repositories. Another respondent proposed taking a survey of faculty who have received grants where data management plans were required to determine their needs and ways to meet those needs. Registering data from external repositories was mentioned once.

- Consult with senior administrators, especially research advisors, deans, and campus leaders to gain support for development of services. (2/4)
- Conduct workshops for graduate students on best practices in data management. (1/4)
- Refine policies for university's data repository, including metadata requirements, format, preservation, quality control, etc. (1/4)
- Survey faculty who have received NSF/NIH grants with data management plan requirements to discover what needs they have in order to tailor future services. (1/4)
- Register data from other repositories in the institutional data repository and cross-link those records with relevant faculty publication in the institutional repository. (1/4)

Respondents repeatedly emphasized that providing data management services fulfilled existing and future needs of their researchers. The need for training is repeated in four out of the five areas of response shown above. The need for campus collaborators shows in all five areas. Through these collaborations, librarians hope to limit the barriers to timely data curation and management services: experienced staff and money.

Conclusions

Fulfilling the missions of university libraries and the institutions of higher education that they serve was a key motivation for librarians to create data curation and management services. Collaborating with the Office of Research, or equivalent, and Information Services is vital for the successful expansion or development of services. The relative ease with which information on data curation and management is being made available online, within one to two clicks, demonstrated thoughtful planning (table 3). Increasing the number of pages with links to multiple resources would increase researcher awareness of these resources. Linking to more outside resources would benefit interdisciplinary researchers (table 4). Determining fair and accurate costs of data storage services, $300 to $500 for universities examined, remains a challenge as universities attempt to project the costs of keeping data sets indefinitely (table 5).

As they work to better determine the current and future needs of their stakeholders, librarians should remain flexible; they should

be aware that today's solutions may be just-in-time services that will be obsolete tomorrow. Librarians should continue to monitor major data initiatives, subject repositories, and industry events such as DataCite, DataONE, Dryad, Figshare, and the O'Reilly Strata conferences (DataCite; DataONE; Dryad 2012; Figshare; O'Reilly Strata 2011). By partnering with government, international, and corporate interests the technological and financial challenges associated with preserving big datasets in sustainable and accessible ways can be overcome.

Declaration of Conflicts of Interest: The author is an employee of the University of Nebraska–Lincoln, one of the institutions analyzed in this research. The author is also a member of the University Libraries Data Curation Committee.

Acknowledgments

Elaine Westbrooks, associate university librarian at the University of Michigan, and Leslie Delserone, assistant professor at the University of Nebraska-Lincoln, consulted in the design of this research.

References

Association of Public and Land-Grant Universities. 2012. "Welcome to APLU." The Association of Public and Land-Grant Universities Website. Available at http://www.aplu.org/page.aspx?pid=203.

Buckmaster, Brooke. 2011. "Grant funds opening of online data storage unit." The Oklahoma Daily Website. Available at http://beta.oudaily.com/news/2011/oct/21/grant-funds-opening-online-data-storage-unit/.

Childers, Scott, and Leslie Delserone. 2011. "The Experienced Viewpoint: Advice on Digital Data Storage and Management Suggestions and Comments from UNL Library faculty Scott Childers and Leslie Delserone." The University of Nebraska–Lincoln Website. Available at http://watercenter.unl.edu/downloads/DigitalStorageAndManagement.pdf.

Clemens, Roxy. 2011. "Suggested Practices: Developing a Data Management Plan for NSF Grant Applications." The Iowa State University Website. Available at http://www.ag.iastate.edu/research/fundingResources/sites/default/files/Developing%20NSF%20Data%20Management%20Plans%20Jan%202011_0.docx

Coalition for Networked Information. 2011. "Data Management Strategies." The CNI (Coalition for Networked Information) Website. Available at http://www.cni.org/topics/digital-curation/data-management-strategies/.

College of Agriculture and Life Sciences. 2012. "Funding Resources." The Iowa State University Website. Available at http://www.ag.iastate.edu/research/fundingResources/.

DataCite. "DataCite." The DataCite Website. Available at http://datacite.org/.

DataONE. "DataONE." The DataONE Website. Available at http://www.dataone.org/.

Deards, Kiyomi. 2012a. "Data Management." The University of Nebraska–Lincoln Libraries Guides Website. Available at http://unl.libguides.com/datamanagement.

———. 2012b. "Definitions." The University of Nebraska–Lincoln Libraries Guides Website. Available at http://unl.libguides.com/aecontent.php?pid=200599&sid=2774497.

———. 2012c. "Documentation & Storage." The University of Nebraska–Lincoln Libraries Guides Website. Available at http://unl.libguides.com/aecontent.php?pid=200599&sid=1676835.

———. 2012d. "FAQ." The University of Nebraska–Lincoln Libraries Guides Website. Available at http://unl.libguides.com/aecontent.php?pid=200599&sid=2774507.

———. 2012e. "Management & Preservation." The University of Nebraska–Lincoln Libraries Guides Website. Available at http://unl.libguides.com/aecontent.php?pid=200599&sid=1676838.

———. 2012f. "Need Help?" The University of Nebraska–Lincoln Libraries Guides Website. Available at http://unl.libguides.com/aecontent.php?pid=200599&sid=1677210.

———. 2012g. "Preserving the Present to Inform the Future: Issues in Data Preservation and Access." Presentation at the Joint Conference of Librarians of Color, Kansas City, MO, September 19-23, 2012. The University of Nebraska–Lincoln Website. Available at http://digitalcommons.unl.edu/library_talks/83/.

Dryad. 2012. "Dryad Digital Repository." The Dryad Website. Available at http://datadryad.org/pages/about.

Duever, Meagan. 2012a. "Data Management." The Kansas State University Website. Available at http://guides.lib.k-state.edu/content.php?pid=298773.

———. 2012b. "Data Management: Archiving/Preservation." The Kansas State University Website. Available at http://guides.lib.k-state.edu/content.php?pid=298773&sid=2512282.

———. 2012c. "Data Management: Data Management Resources." The Kansas State University Website. Available at http://guides.lib.k-state.edu/content.php?pid=298773&sid=2459535.

———. 2012d. "Data Management: Metadata." The Kansas State University Website. Available at http://guides.lib.k-state.edu/content.php?pid=298773&sid=2480281.

———. 2012e. "Data Management: Organizing Your Data." The Kansas State University Website. Available at http://guides.lib.k-state.edu/content.php?pid=298773&sid=2464009.

———. 2012f. "Data Management: Sharing Your Data." The Kansas State University Website. Available at http://guides.lib.k-state.edu/content.php?pid=298773&sid=2512282.

Figshare. "Figshare." The Figshare Website. Available at http://figshare.com/.

Iowa State University. a. "National Science Foundation Data Management Plan Requirements." The Iowa State University Website. Available at http://www.ospa.iastate.edu/docs/NSF-Data-Management-Plan.pdf.

———. b. "Creating Data Management Plans for NSF Proposals: Template." The Iowa State University Website. Available at http://www.ospa.iastate.edu/docs/CreatingDataManagementPlansforNSFProposals.pdf.

———. 2012. "Office of Sponsored Programs Administration." The Iowa State University Website. Available at http://ospa.iastate.edu/.

Kansas State University. 2012a. "K-State Libraries: Divisions and Departments: Faculty and Graduate Services." The Kansas State University Website. Available at http://www.lib.k-state.edu/depts/.

———. 2012b. "K-State Libraries: K-State Data Services." The Kansas State University Website. Available at http://data.k-state.edu/.

Kansas State University Office of Research and Sponsored Programs. 2012a. "Classes, Workshops, and Training by Date." The Kansas State University Website. Available at http://www.k-state.edu/research/resources/calendar/index.htm.

———. 2012b. "Data Management Plan (DMP) Information Sessions." The Kansas State University Website. Available at http://www.k-state.edu/research/resources/calendar/2012-05-03DataMgmtPlan%20Agenda.pdf.

———. 2012c. "NIH Data Management Plans." The Kansas State University Website. Available at http://www.k-state.edu/research/resources/writing/Data%20Mgmt/NIH%20Data%20Mgmt/index.htm.

———. 2012d. "NSF Data Management Plans." The Kansas State University Website. Available at http://www.k-state.edu/research/resources/writing/Data%20Mgmt/NSF%20Data%20Mgmt/index.htm.

———. 2012e. "NSF's Data Management Plan Requirements: Kansas State University Guidance for Sharing and Archiving Data." The Kansas State University Website. Available at http://www.k-state.edu/research/resources/writing/Data%20Mgmt/NSF%20Data%20Mgmt/NSF%20Data%20Management%20Plan%205_2_12.docx.

———. 2012f. "Proposal Writing Resources." The Kansas State University Website. Available at http://www.k-state.edu/research/resources/writing/index.htm.

Neeman, Henry, and Patrick Calhoun. 2012. "The Oklahoma PetaStore: Large Scale Storage for Research Data Archiving and Management." The Great Plains Network Website. Available at http://www.greatplains.net/download/attachments/3280974/Henry+Neeman+PetaStore.pdf.

Oklahoma State University. a. "Helpful Links." The Oklahoma State University Website. Available at http://urs.okstate.edu/index.php/helpfullinks.

———. b. "NSF Data Management Plans—Suggested Structure (note 2-page limit)." The Oklahoma State University Website. Available at http://ucpd.okstate.edu/resources/nsfdmp/dmpstructure.

Oklahoma State University. 2011. "Special Events." The Oklahoma State University Website. Was available at http://www.ceatresearch.okstate.edu/pgs_spec_events.aspx (*note, this URL no longer exists*).

Oklahoma State University. 2012a. "In the Know." The Oklahoma State University Website. Was available at http://www.ceatresearch.okstate.edu/pgs_in_the_know.aspx (*note, this URL no longer exists*).

———. 2012b. "NSF Data Management Plans." The Oklahoma State University Website. Available at http://ucpd.okstate.edu/resources/nsfdmp.

———. 2012c. "NSF Data Management and Sharing Plans." The Oklahoma State University Website. Available at http://urs.okstate.edu/docs/NSFDataManageandSharingPlans.pdf.

———. 2012d. "Resources." The Oklahoma State University Website. Available at http://ucpd.okstate.edu/resources.

Oklahoma State University Office of Proposal Development. 2011. "Data Management Plans." The Oklahoma State University Website. Available at http://urs.okstate.edu/images/stories/DMP4-7-11OPD.pdf.

O'Reilly Strata. 2011. "O'Reilly Strata Conference: Making Data Work." The O'Reilly Strata Conference Website. Available at http://strataconf.com/.

Smithsonian Institution. 2012. "Public and Land-Grant Universities and the USDA at 150." The Smithsonian Institution Website. Available at http://www.festival.si.edu/2012/campus_and_community/.

tfedderson2. 2011. "Data Management Assist Available to Researchers." The University of Nebraska–Lincoln Website. Available at http://scarlet.unl.edu/?p=10242.

United States Department of Education. "12/2012 Accreditation Files." The United States Department of Education Website. Available at http://ope.ed.gov/accreditation/GetDownloadFile.aspx.

University of Nebraska–Lincoln. 2012a. "Data Curation—Academic Activities Brown Bag." The University of Nebraska–Lincoln Website. Available at http://mediahub.unl.edu/media/2488.

———. 2012b. "Data Curation FAQ Session and Discussion." The University of Nebraska–Lincoln Website. Available at http://mediahub.unl.edu/media/2472.

———. 2012c. "IANR Landgrant Legacy Celebration." The University of Nebraska–Lincoln Website. Available at http://libraries.unl.edu/landgrant.

———. 2012d. "Libraries: Data Management." The University of Nebraska–Lincoln Website. Available at http://libraries.unl.edu/data-management.

———. 2012e. "Libraries Offers Assist with Data Management Plans." The University of Nebraska–Lincoln Website. Available at http://newsroom.unl.edu/announce/todayatunl/577/3342.

———. 2012f. "Office of Research & Economic Development." The University of Nebraska–Lincoln Website. Available at http://research.unl.edu.

———. 2012g. "Office of Research & Economic Development: Faculty Resources." The University of Nebraska–Lincoln Website. Available at http://research.unl.edu/facultyresources/.

———. 2012h. "Office of Research & Economic Development: Proposal Development." The University of Nebraska–Lincoln Website. Available at http://research.unl.edu/proposaldevelopment/.

———. 2012i. "Office of Research & Economic Development: Proposal Development: NSF Data Management Plan Resources." The University of Nebraska–Lincoln Website. Available at http://research.unl.edu/proposaldevelopment/NSFresources.shtml.

———. 2012j. "University Libraries Strategic Plan 2012-2013." The University of Nebraska–Lincoln Website. Available at http://libraries.unl.edu/docs/120403_Srategic_plan_2012-13.pdf.

———. 2012k. "University of Nebraska–Lincoln Libraries." The University of Nebraska–Lincoln Website. Available at http://libraries.unl.edu/.

———. 2012m. "UNL Data Repository." The University of Nebraska–Lincoln Website. Available at https://dataregistry.unl.edu/.

———. 2012n. "What's New." The University of Nebraska–Lincoln Website. Was available at http://libraries.unl.edu/news (*note, this URL no longer exists*).

University of Oklahoma. 2012a. "OU Supercomputing Center for Education & Research." The University of Oklahoma Website. Available at http://www.oscer.ou.edu.

———. 2012b. "PetaStore: OU Supercomputing Center for Education & Research: OSCER PetaStore Policy and Procedures." The University of Oklahoma Website. Available at http://www.oscer.ou.edu/petastore.php.

UNL Data Repository. 2011. "Create Account." The University of Nebraska–Lincoln Website. Available at https://dataregistry.unl.edu/account.jsp.

UNL Office of Research. 2011. "Resources for NSF Data Mangement Plans." The University of Nebraska–Lincoln Website. Available at http://research.unl.edu/researchnews/April2011/story.php?ID=rr-nsf.

Washington State University Extension. 2009. "What is a Land-Grant College?" The Washington State University Extension Website. Available at http://ext.wsu.edu/documents/landgrant.pdf.

Westbrooks, Elaine. 2011. "Archival Data Management for Research." The Great Plains Network Website. Available at www.greatplains.net/download/attachments/590157/Data+Management.pptx.

Westbrooks, Elaine, and Kathryn Notter. 2011. "Libraries for Seamless Data Archiving." The Prezi Website. Available at http://prezi.com/j3q3fzxj66-w/the-university-of-nebraska-lincoln-data-repository-partnerships-between-it-and-libraries-for-seamless-data-archiving/.

APPENDIX: E-mail Interview Script

Thank you for your willingness to help. Attached is the informed consent form for performing this self-interview. Please read the informed consent form and save a copy for your records.

Before you begin I'd like to remind you that you are free to skip any question you are not comfortable answering. Simply leave the question blank and move on to the next one you are comfortable answering. Because you are self-interviewing, I have included sub-questions to think about on questions 1, 3, and 6. If you have any questions about this self-interview, please feel free to e-mail me at kdeards2@unl.edu.

1. What services do you offer in regards to Data Management and Curation?
   - Why were these services developed?
   - Who led their development?

2. How are the library's services tied into the campus' goals, policies, and needs, regarding data management?

3. How do you collaborate with other campus units in regards to data curation?
   - Why were these collaborations created?

4. What has been successful and why?

5. What has not been successful and why?

6. What do you see as the next steps to expanding or refining existing data management and curation services?
   - What do you see as barriers to achieving these steps?
   - What do you see as advantages to taking these steps?

How important do you feel that Data Management and Curation will be in the future?

Responses to Data Management Requirements at the National Scale

Chris Jordan, Maria Esteva, David Walling, Tomilsav Urban, and Sivakumar Kulasekaran

Introduction

The introduction of data management plan requirements by the National Science Foundation (NSF) has drawn new attention to the need for data management infrastructure, both in terms of hardware and in terms of human and policy support, within the realm of academic research. The National Institutes of Health (NIH) has required specific data sharing plans for several years, as do other federal funding agencies (NSF n.d.). For large collaborative research projects such as the data processing associated with the Large Hadron Collider, data management plans have been an internal requirement for many years, and significant effort may go into preparing and testing data management mechanisms even before major data generation gets under way.

Individual institutions have responded to these requirements by providing researchers with consulting services, web pages, and template data management plans, often with significant participation from research libraries and library staff. Regional and national-scale institutions and networks have also recognized the need to support data management planning and execution, and have worked to develop effective mechanisms to address the challenges that researchers face. As a result, approaches to data management vary in structure, level of financial support, length of retention, and openness for data sharing and collaboration; in addition, the costs to researchers vary, ranging from the provision of infrastructure at no cost to the use of commercial providers with the costs passed directly to the researchers. We present some of the issues driving the need for data management support beyond individual institutions, discuss some current approaches at the regional and national scale, and explore the potential interaction of these approaches to provide a more

complete fabric of support for research data management for researchers at all scales, from individual researchers working at small institutions to large multi-institutional projects.

Although it is now required that researchers submit formal data management plans alongside requests for funding from NSF and other funding agencies, data management has always been a requirement for research using digital data, and the issues related to data management have existed for both local and national-scale institutions for many years. The value of the new requirements is that they give increased visibility to an important aspect of contemporary research processes and create a sense of urgency for individual researchers to more directly grapple with an issue that has been in the background for the last decade or more, as attested by numerous reports commissioned by the U.S. government (National Science and Technology Council 2009) and academic organizations (Association of American Medical Colleges 2011).

In this paper, we will address primarily the issues related to supporting the development and implementation of data management plans for research, as opposed to the process of writing data management plans. The former are of greatest concern, as it is difficult to write a plan without a feasible set of resources to support it, and providing such resources requires considerable investment and expertise. Thus, when considering data management requirements and the resources necessary to support them, we focus on this broader perspective of research needs.

We believe that it is particularly informative to focus on national-scale projects, as they are well positioned both to achieve a broad understanding of the data management needs of the communities they serve and to participate in the development of a consensus about the best practices for the planning and implementation of data management programs for research. Some of these projects are struggling to implement storage solutions; some are designing architectures to support distributed storage; and some are developing best practices and software services to address real or perceived challenges for data management. Regardless, because of their size and the size of the communities that they serve, their successes and failures will become a part of the ongoing discussion about appropriate short- and long-term data management practices, and the infrastructure needed to support them.

Research Data Management Requirements

There is an inherent need for data management in the conduct of contemporary research, but in addition, national funding agencies have established data management requirements in recent years. Often, funders require only that an explicitly titled data management plan be submitted alongside grant requests; the plan may consist entirely of an argument that a given research project will generate no data requiring management. However, data management plans will be subject to the same peer review processes that the NSF and

other government agencies use to evaluate proposals. Consequently, research communities are likely to eventually develop disciplinary norms that will govern the expectations of data management plans. That is why national-scale responses to these requirements are of strong interest. On the one hand, it is now common knowledge that data management is an important part of the research planning and funding selection process. On the other hand, there are not yet widely accepted norms for what constitutes an appropriate data management plan; the understanding of what is required in the conduct of research is only developing.

Planning and implementing the handling of digital data in an academic context is an inherently multidisciplinary process. Research involving digital data frequently involves computer and information scientists and engineers, and these same areas of expertise are increasingly being called upon to support a variety of digital data workflows, such as metadata definition and generation; construction of customized analysis workflows and the interfaces to those workflows; and the use of computational and storage infrastructure to support storage, collaborative access, and preservation of digital data. The need for multidisciplinary involvement common to many large and small research projects often requires cross-institutional collaboration or the recruitment of expertise from outside a given researcher's usual network of collaborators, which is one of many forces pushing data management concerns to the regional and national level. Large-scale projects providing advanced computational and storage infrastructure are being called on to support an ever-widening circle of research projects.

An important issue that receives relatively little attention in discussions of data management planning is the basic need for effective and, in many cases, large-scale infrastructure on which to store and analyze data. A related but less urgent need is for advanced computational resources and techniques to perform needed analysis or create needed visual representations of complex data sets. Because data management must begin with data, and because those data must be stored somewhere, basic storage capacity is an absolute minimum requirement that can be surprisingly difficult for researchers to satisfy. Individual institutions have significantly increased their own technological capabilities in recent decades, but limited budgets for all departments and the extreme specialization of much advanced technology has made it ever more difficult for any one institution to provide the infrastructure and expertise required to support diverse research needs. In particular, the issues of data management involve not just the scale of the required infrastructure, but also the diversity of the digital data services needed to support contemporary research activities. Any number of commercial providers, as well as most campus-level information technology departments, now provide basic file storage, but many current projects require a specific database implementation, including novel structured data storage techniques such as the NoSQL approaches taken by MongoDB[1] and

[1] http://www.mongodb.org

similar tools. Web-based services are ubiquitous and include many discipline-specific web applications and even format-specific implementations, such as those for functional magnetic resonance imaging data or for various DNA sequencing applications.[2] Web-based applications themselves typically rely on a variety of underlying software tools and storage engines, including various web frameworks, scripting languages, and databases. Data management for multiple research projects requires infrastructure and expertise in support of all these types of applications and more, and institutions with limited funding are finding it increasingly difficult to handle the data management needs of their research communities. Therefore, researchers increasingly need to use regional and national-scale partnerships and institutions to find the support required to execute their research data management plans.

Review of National-Scale Cyberinfrastructure Projects

There are a variety of projects and institutions that are chartered explicitly to support research through computational and storage infrastructure, or have some engagement with digital data management concerns. These projects include the longstanding NSF cyberinfrastructure initiatives (Atkins et al. 2003), the most relevant of which are the eXtreme Science and Engineering Discovery Environment, the projects funded under the DataNet program, and new discipline-specific initiatives (e.g., iDigBio and iPlant); they also include regional consortia or statewide university systems such as the university systems of California and Texas. There are a few more recent projects that have been formed specifically to address issues of data management and preservation in multi-institutional ways, including HathiTrust[3] and the Digital Preservation Network,[4] among others. Finally, there are partnerships in which an academic or research institution has established a front-end interface or a special arrangement with a commercial infrastructure provider in ways that are intended to facilitate research data management or general access to storage resources for any data type. We will discuss several of these projects to provide some sense of the contrasting approaches to supporting data management at the national scale, including the areas of overlap and meaningful gaps in the overall infrastructure.

The eXtreme Science and Engineering Discovery Environment

The NSF-funded eXtreme Science and Engineering Discovery Environment (XSEDE) links staff and resources at supercomputing centers around the United States to provide advanced digital services in support of research. A successor to the TeraGrid program, XSEDE

[2] See, for example, http://xnat.org.

[3] http://www.hathitrust.org.

[4] http://d-p-n.org

represents a continuation of NSF's longstanding support of high-end infrastructure for research computing. Historically, the programs have focused on high-performance computing and visualization, but with XSEDE, the focus has significantly broadened to reflect the increasing importance of digital data and data services to the research enterprise. As such, the XSEDE project has received requests both from funding agencies and from users for enhanced support for data management planning and execution. This support takes the form of resources for storage of digital data, services supporting access and collaboration using digital data, and training on data management planning and execution practices.

Resources provided by XSEDE for digital data consist primarily of petabyte-scale tape archive systems, which have been provided by the supercomputing centers for decades, but have been directly associated for the most part with computation taking place on the resources provided by those centers. Increasing demands for online storage and new access mechanisms have led to the provision of several wide-area file systems accessible from multiple resources (Andrews, Jordan, and Lederer 2006), web interfaces for managing data,[5] and plans for new services to manage the replication of data across multiple systems. One of the most significant obstacles to the use of XSEDE for data management planning is the fact that resources are allocated through an annual peer review process; as a result, researchers cannot write a data management plan for a multiyear submission to NSF with certainty that XSEDE resources will be available to them for the entirety of the project. A variety of solutions have been proposed to resolve this conundrum, including directly associating NSF grants of funding with XSEDE grants of storage resources, allocating storage resources on a multiyear basis, and providing long-lived "community" allocations that can be used by many researchers from various institutions within a specific disciplinary community.

The more general problem is that XSEDE exists to support active research for short periods of time, while data management tends to be a long-term activity. XSEDE is not expected to provide support for long-term management of data or the integration of data into larger "reference" data collections. Rather, the intention is that other institutions with a more long-term mission or a specifically disciplinary focus will handle the ongoing issues of data management, while XSEDE focuses on the shorter-term issues of helping users store and utilize data for individual research initiatives. However, this does leave responsibility for identifying such long-term resources and transitioning data up to the researchers themselves. If XSEDE chooses not to provide for long-term storage, one obvious area of potential improvement for XSEDE is to facilitate the connection of researchers and data collections to appropriate reference repositories for long-term storage and to automate the migration of data between XSEDE resources and these repositories.

[5] See, for example, XSEDE SHARE web-based file-sharing service at http://share.xsede.org/ShareService.

XSEDE offers significant training and user support, including in-person and webcast training events open to anyone, as well as systems for addressing specific issues posed by the user community. Some training events have focused on general development and execution of data management plans, while others have addressed specific topics such as the use of structured data and metadata. Webcast training events are recorded and made available online for viewing by any researcher or interested individual. In addition, an initiative created in the TeraGrid and expanded in XSEDE has established a pool of experts in various technical areas who can be requested and allocated through a peer-review process, much as hardware resources are allocated. If such a request is granted, a research team receives dedicated support from one or more members of the XSEDE team, including support for developing and executing data management plans using both XSEDE resources and other resources available to the researchers. This support, which can extend for up to a year, can provide significant aid to researchers in dealing with data management challenges that may be new to them and for which appropriate advice and support might not otherwise be available.

Although XSEDE provides both a mechanism to help researchers find appropriate data management resources and growing support for storage and sharing of research data, significant areas of research data management are not well supported through XSEDE. Because XSEDE offers access primarily to resources at member centers, along with services for managing data within and between those resources, it lacks provisions for many of the specialized, persistent data services that are required for some research activities. These services include database support, support for customized web services and web collaboration tools, and support for open data sharing. Member centers support some of these services, but in those cases, they are available only to users from the local institutions and not to those at the national scale. However, because the number of software and hardware tools that may be needed for various research data management tasks is so large, it is unlikely that even a national infrastructure such as XSEDE would be able to support anywhere close to the full range of necessary tools. Instead, XSEDE must focus on the requirements of the largest communities of users and rely on its member centers or local institutions to provide the less widely used infrastructure and expertise.

The Digital Preservation Network

A more recent development that promises to offer support for long-term data management is the Digital Preservation Network (DPN). Funded by a consortium of research universities, DPN will provide policy, legal, and network infrastructure to support long-term preservation of important data, including scholarly publications and the research data on which they are based. Initially, DPN will rely on a network of five independent preservation infrastructures, which will be connected to each other and will collaborate to develop a technical architecture that supports submission of data from a member

institution, replication of those data to multiple locations, and, when necessary, the retrieval of data from one of the replica locations.

This network is meant to function in accordance with a set of policies and legal procedures established to ensure not only that data are protected but that the legal ownership of data and the rights associated with it are also preserved in such a way that data remain available over time, even in the event of institutional failure. Therefore, rather than facilitating ongoing access to or collaboration in using data, DPN will focus on preserving the data in a relatively static format. In other words, it is meant to provide a solution for the deposit of data after active research using those data has concluded. DPN represents an important contribution to the challenge of data management—one that no individual institution could provide. As the project evolves, it will be necessary to develop mechanisms to access data from local repositories and national projects (e.g., XSEDE) that maintain the data in the short term and transfer them into the long-term preservation infrastructure, along with appropriate ownership and rights information.

Institutions or projects participating in DPN include several that are themselves multi-institutional or regional in nature, or both, such as HathiTrust, Chronopolis,[6] and the University of Texas (UT) Research Data Repository. The difficulties of constructing a long-term, national-scale solution from a number of regional-scale or multi-institutional initiatives point to the complexity of the challenges involved and the solutions required, as well as to the scale and diversity of the institutions that are providing components of the data management framework that is developing in the United States.

Both the DPN and the XSEDE projects are essentially federations of existing resources; they provide services in addition to those resources, but do not, with few exceptions, provide hardware infrastructure directly. This is a relatively common characteristic of national-scale efforts to support data management, as funds are limited and are directed primarily toward the development and deployment of services, itself a complex task given the nature of the technology and the number and diversity of the underlying resources—as opposed to the direct provision of resources. However, both DPN and XSEDE perform an important function in connecting researchers who have data management needs with available resources on which to store and manage data.

Integrated Digitized Biocollections

Like the XSEDE project, the DPN project is general in its focus; that is, both projects provide cyberinfrastructure to support research data management regardless of the discipline or type of data involved in the research. In contrast, some other projects develop national resources to provide data management, access, and preservation for a specific data type or in a specific discipline. One example, the Integrated Digitized Biocollections (iDigBio) project, focuses on

[6] See http://chronopolis.sdsc.edu.

digitized materials from collections of individual physical specimens from the natural world, traditionally referred to as natural history collections. Although they may engage many individuals with their own research programs, these collections do not typically focus on research itself but on the provision of data used in research, data that are increasingly converted into digital form and accessed over a network by researchers around the world. iDigBio is developing and implementing an infrastructure for storage, search, and retrieval of digitized collections data, which will allow institutions throughout the United States to register digitized materials used in the course of research. As part of its Advancing Digitization for Biological Collections initiative, the NSF is providing funds both for the digitization programs generating the data and for the iDigBio project itself. The iDigBio project will be largely responsible for the data management processes in collaboration with more specific thematic networks that will supervise the digitization processes.

The iDigBio model is significantly different from the XSEDE and the DPN projects in that the data management will be undertaken and supported primarily by the national-scale projects using well-defined processes, and data will be provided by the individuals and institutions participating in the project, as well as data that are generated independently of the iDigBio project itself. The data that will eventually be accessed and managed by iDigBio will need to be stored and made accessible for an indeterminate period of time (i.e., forever), and thus to a certain extent, digital preservation will eventually be a critical concern. At this stage of the project, however, the focus is much more on data generation, indexing, and access than on long-term stewardship.

The iPlant Collaborative

An NSF-funded initiative focusing on cyberinfrastructure for plant science, the iPlant Collaborative represents another discipline-centered national resource with a significant data management component.[7] iPlant provides general cyberinfrastructure, including software, computational capacity, web interfaces to automate workflows, and data storage and management tools, to individual plant scientists regardless of their home institution. It also provides customized access to backend infrastructure to very large collaborative projects that have specialized and intensive data management needs, such as the One Thousand Plant Transcriptome project, for which iPlant provides data storage and access for the many distributed researchers involved. The data management components of iPlant have been among the most widely adopted and successful of its services, with hundreds of researchers from institutions across the United States using the web-based interfaces to store, share, and collaborate on projects using hundreds of terabytes of research data. This success is due partly to iPlant's provision of automated workflows to perform many common data type-specific analyses and management

[7] See http://www.iplantcollaborative.org.

tasks, and partly to the fact that researchers can share workflows and data. Thus, the iPlant infrastructure enables not just the execution of data management tasks, but also the sharing of best practices along with the research data. It is expected that the quantity of data stored within the iPlant infrastructure will continue to grow and that the iPlant Collaborative will continue to add new functionality (in particular, enhanced storage of metadata in general and support for a broader set of provenance metadata) to further facilitate sharing of data between researchers.

In comparative terms, iPlant represents a kind of middle ground between the XSEDE and the DPN projects and the iDigBio approach in the level of centralization and the diversity of life cycle stages and communities that it supports. Although iPlant's data management resources are mostly under the control of the project, the data themselves are almost entirely provided by external researchers; the project exerts very little control over the data submitted to and used within the infrastructure. With iDigBio, the project is more closely involved in data generation and must impose a certain degree of uniformity on data structure and contents in order to facilitate cross-collection indexing and search functions. Also, the development and promotion of best practices is a more explicit component of the iDigBio mission. iPlant does not include the long-term preservation of research data as a part of its mission, and much of the data managed by researchers within the infrastructure are stored there for only relatively short periods of time while they are in the active research stage of the data life cycle. How iPlant will handle data that have more significant value, but may not have a natural home for long-term preservation is an issue that remains to be addressed.

DataNet

The NSF's Sustainable Digital Data Preservation and Access Network Partners program, referred to as DataNet, was widely expected to provide significant advances in support for data management because of the generous funding available and the ambitious set of goals expressed in the program solicitation. As indicated in its title, one goal of the DataNet program was to establish a nationwide network of program awardees that would be capable of assessing and responding to the challenges of data management as they arose; in practice, however, the various projects have operated somewhat independently and have not acted as a single coordinated institution. Therefore, we have not attempted to treat DataNet as a model for direct comparison for projects such as XSEDE or DPN. Better models for comparison are the European Preservation and Long-term Access through Networked Services (PLANETS) project and, to a lesser extent, the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP), which have focused on a variety of challenges related to research data management, preservation, and access.

Most DataNet awardees, like those of the PLANETS and NDIIPP programs, have addressed various aspects of data management

challenges through software components, development of standards for digital object handling, and network services rather than directly providing resources or working with researchers to deal with data management challenges. These components all provide necessary pieces of the overall data management puzzle, but do not directly address the challenges of research data management faced by individual researchers. Instead, they may form components of solutions offered by national projects such as DPN, XSEDE, or iDigBio. We discuss the DataNet Federation Consortium (DFC)[8] as an example of the kinds of activities undertaken under the DataNet umbrella.

The DFC includes six large-scale, collaborative research projects with a team of software developers, primarily at the Renaissance Computing Institute at the University of North Carolina. They are working to improve an existing open source system for data management, the Integrated Rule-Oriented Data System (iRODS) software.[9] The iRODS software is intended to support the management of data in large federations of storage systems, or "data grids," with a focus on the development and implementation of rule-based systems. The research projects, which include the iPlant Collaborative along with projects from the earth sciences, the social sciences, the humanities, and the physical sciences, provide the requirements of their data management workflows, and the core DFC team implements the software tools and the policies required to support these specific workflows. In addition, the DFC project tools and policies are expected to work in more general contexts. As with other DataNet partners, the DFC project does not directly provide infrastructure for the storage of research data, but it does provide a critical component in the research data management stack. The engagement of the project with existing large-scale research collaborations allows the components created in the course of the project to evolve so that they become best practices as the supported communities grow more sophisticated in their data management activities.

A final national-scale response to growing data management requirements that should be mentioned here is academic and research project partnerships with cloud storage providers that have built out national infrastructures for storing and replicating data. Perhaps the best known of these providers is DuraCloud,[10] which uses storage from Amazon S3 and other commercial providers and passes the costs on to the data providers; this effort is now well documented and seems to work well for the specific digital library application space to which it is targeted. A more general solution following the same pattern is the Internet2–InCommon partnership with Box.net,[11] a commercial cloud storage provider, that allows Internet2 member institutions to access the service at a reduced cost, using the Internet2 network. This partnership is not aimed exclusively at research data

[8] See http://datafed.org.

[9] See http://www.irods.org.

[10] See http://www.duracloud.org.

[11] The partnership, NET+Box, is described at http://www.internet2.edu/netplus/box/index.html.

storage, and commercial services such as Box.net tend to be designed more to support the sharing of office documents, images, and audio files that are relatively small, as well as limited use cases, rather than to deal with the breadth and depth of research data management requirements. This is a fairly new service, and the authors do not know of any specific experiences with its use in support of research data storage. However, this kind of partnership with commercial infrastructure providers does represent another potential model for the national provision of resources for data management, and the broad participation in Internet2 and its past success in providing network infrastructure supporting research implies that this partnership has some promise. It will be important to monitor the use of this service as an example of a national partnership between an academic/research-focused organization and a commercial information technology service provider, as it is likely that other, similar partnerships will be formed in support of various aspects of research data management and analysis in the years to come.

Relationship of National to Regional Responses to Data Management Requirements

All the projects discussed in the preceding section either rely directly on or interact to some extent with infrastructures developed to address more local concerns, particularly those hosted by individual institutions with an explicit goal of serving a regional or national community. For example, both the San Diego Supercomputing Center (SDSC) at the University of California, San Diego, and the Texas Advanced Computing Center (TACC) at the UT at Austin are participants in the XSEDE project and have large-scale storage infrastructures aimed at broad regional or national communities. Both also provide part of the underlying infrastructure for the DPN initiative, through the Chronopolis project in the case of the SDSC and through the UT Library's Digital Repository in the case of the TACC. There are many other examples, but as both these institutions are part of very large university systems, they provide useful models of the way in which regional-scale or multi-institutional resources may provide a bridge between smaller individual institutions and national-scale resources. Such a link is particularly important for individual researchers and smaller projects that may struggle to compete with larger and more established projects for the support of national institutions.

SDSC has a long history of support for data collections and data management, and recently announced the availability of a new "Cloud Storage" resource, both for researchers at the University of California and for other institutions, on a cost-recovery basis.[12] The resource will eventually include geographic replication. It has a raw

[12] Information on the SDSC Cloud Storage System is available at https://cloud.sdsc.edu.

capacity of 4 petabytes, and it has no restrictions on access besides the ability and willingness of the user to pay. In many ways, this resource is analogous to the Box.net service or the underlying storage provided by DuraCloud, and in fact, the DuraCloud system will soon support the use of the SDSC's Cloud Storage instead of the use of a commercial storage provider. Although there is some interest in this model as an emerging pattern by which to provide researchers nationwide with access to basic storage in support of data management, this resource will also be used in support of large-scale projects such as Chronopolis, DPN, and other well organized, multidisciplinary collaborative research efforts. The cloud storage model is also designed to be integrated with existing or developing web-based infrastructures for data management and collaboration, and it is intended to allow researchers to execute data management plans and to meet growing needs for digital data with a minimum of complexity and no need to wait for cumbersome peer review processes.

The UT system, representing nine academic campuses and six medical campuses distributed throughout the state of Texas, provides a contrasting approach with its Research Cyberinfrastructure (UTRC),[13] a systemwide initiative encompassing high-performance computing, a 10-gigabit research network, and a 5-petabyte research data repository. These components were deployed following a broad consultation with the research community across the UT campuses, in which research data storage was identified as a critical need for all disciplines, particularly the life sciences. As a result, the system funded the research data repository, hosted by the TACC as a resource comparable to the SDSC Cloud Storage, which provides up to 5 terabytes of geographically replicated storage to any principal investigator at a UT campus at no cost. The system also funded a research network to link all 15 UT campuses to each other and to the research data storage at 10 gigabit per second speeds, with a further requirement that research facilities within the campuses have direct access to the full network capacity. Additional efforts have been undertaken to provide training in the use of these resources for data management and to support collaborative projects involving multiple institutions, inside and outside of the UT system. Contributions from research projects that need larger allocations of storage will fund further expansion of the system, and consultation with the user community on needs for data management resources and services will drive future developments.

Besides the fact that the TACC storage resource is not based on a cloud storage model, the other major difference between the SDSC and the TACC resource is that the TACC resource is available at small scales to all researchers at all UT campuses free of charge. There are also similarities between the two initiatives. For example, like the SDSC resource, the UT research data repository can be used by both individuals and projects operating at scales up to national collaborations, and data management and sharing activities are

[13] See http://www.utsystem.edu/research-cyberinfrastructure/homepage.htm.

supported throughout the research data life cycle. Also, both these resources suggest the importance of robust, high-performance, and high-capacity storage infrastructure without disciplinary restrictions or other conditions on access as the foundation for the growth of research data management capacity and the development of best practices. Another characteristic in common is that both are connected to ongoing collaborations between cyberinfrastructure providers and campus research libraries, in which the cyberinfrastructure providers support digital library efforts through backend storage and other technology, while library staff provide interfaces, metadata, and other curation expertise. In this way, institutions work together to develop and implement long-term preservation plans. This developing pattern of collaboration, which is also seen at the national scale in the DPN project and elsewhere, has both promise and importance in providing comprehensive solutions for research data management.

Conclusions: A Developing Ecosystem of Data Management Services

The national and regional institutions and projects that have been discussed here cover a broad terrain, both in terms of the types of research data management that they support and the stages of the research data life cycle that they are intended to address; their activities range from active data management to long-term preservation and include everything from single-investigator work to large-scale collaborations. However, researchers still have unmet needs at the national scale, particularly with regard to the transitions between stages of the research lifecycle, as well as at the level of small-scale but complex needs. Although some of these issues may be addressed by local institutions, others require broader, systemic solutions.

Among the issues to be resolved is the question of how researchers can identify the appropriate resources for their current and future projects. Also, how can data management and support responsibilities best be transitioned between institutions as the research effort grows or as researchers' careers change? There are not yet policies, agreements, or facilities to help researchers manage their data as the scale of a research effort or data collection increases or as local researchers join the large, distributed collaborations that are becoming more and more common in the contemporary research environment. Similar issues arise when researchers change institutions, and thus their research data management support infrastructure changes; in many cases large amounts of data need to be transferred. Many significant issues can arise when collections of data amassed over many years at great effort and expense must be migrated between institutional resources, often in short periods of time. As the products of scholarly effort are more frequently born digital, these issues will become both more serious and more common. Local information science institutions, particularly libraries, can help to mitigate the problem of matching resources, human or machine, to individual researchers or projects, but addressing the flow of research data

management over the course of a scholar's career, or as circumstances change, remains a significant issue.

Related issues include legal rights to data and security concerns around data, particularly human subjects data that may be subject to legislative or institutional restrictions on access. None of the national or regional projects that we have discussed directly address the issues of protected data. Although legislation covering, for example, protected health information is extensive and institutions have developed robust internal systems to ensure appropriate handling of research subjects, such legislation and internal controls typically do not provide sufficient detail on implementation to serve as a guide for large-scale management and collaboration using such data. Like the questions of institutional transitions and changes of scale, these issues should be addressed in a multi-institutional fashion, as a collection of individual solutions will inevitably create a balkanized landscape of data collections inside walls with differing rules and protections. Even a common data management practice such as the embargo of related data until a publication date has very little infrastructure implemented to support the automation of such practices, particularly when data are dispersed across multiple systems. It is to be hoped that policy-oriented efforts like the DFC project will begin to address the automation of such practices in the future.

The one certainty regarding research data management, both in terms of planning and execution, is that the needs will continue to grow at an exponential pace in years to come. The number of data objects requiring management; the size of those objects; their past and future relevance; and the communities that wish to collaborate on creating, analyzing, and searching those objects will all increase, as will the complexity of the human and technical infrastructure required to support data management activities. Therefore, it is extremely unlikely that any one institution, no matter what the scale, will be able to support even a majority of the community needs. We have documented a variety of projects and institutions with differing approaches to supporting research data management; they have included substantial human expertise, training activities, discipline- and project-specific interfaces, storage infrastructure, and allocation/cost models. Not all of these approaches are likely to be successful over the long term, and projects are likely to pioneer new models in years to come. We view this as a positive. As with the academic enterprise itself, it is likely that many different solutions will work for different individuals and groups, and that disciplines may choose to organize their data management practices in different ways. It will take a great deal of experimentation and time for such organic growth of the ecosystem to occur.

The ecosystem is a natural analogy to use for the myriad institutions and scales at which research data management is being practiced and supported. The task is itself multifaceted, with information science, policy, legal, and technical components, to name just a few. The development of many different projects with overlapping and contrasting goals will advance the practice of data management by

addressing different challenges at different scales. As previously noted, it is unlikely that a single initiative will be able to solve all the challenges of research data management, even for a single discipline. Thus, the organic, interlocking efforts of many different projects at different scales can be viewed as an absolute requirement for the natural development of accepted practices. As these practices are developed and shared, information technology departments will improve their ability to identify and solve problems at the regional and national scale. Promoting awareness of the broader landscape at the local level is crucial, as this organic ecosystem can benefit only those who know that it is available to them and are willing to use it.

Developing these abilities will require a great deal of interdisciplinary and interinstitutional communication and collaboration. In particular, efforts at the University of Texas, the University of California, and the DPN project have begun to show the value of collaboration between libraries and information scientists, and between information technology departments and advanced technology centers, to improve the flow of communication and promote more holistic approaches to data management. It also seems likely that the discipline of digital data curation will become a better defined career path for information science students in the future, with specializations developing in biological data, engineering data, and so on. National-scale organizations with a significant investment in promoting good curation practices and the integration of digital data curation as a well-defined component of research data management have not yet appeared, but it is to be hoped either that such organizations will be created or that existing organizations will adopt a more vocal advocacy role regarding the importance of sound data curation practices to the global research enterprise.

The prospects for future national-scale data management infrastructure currently look bright, with the NSF's Data Infrastructure Building Blocks[14] and BigData[15] programs acting as follow-up programs to the DataNet initiative. These programs, along with others announced recently as part of a government-wide big data initiative, are likely to provide significant funding for a variety of projects supporting research data management. It will be especially helpful if these programs explicitly recognize the need for projects at multiple scales, from exploratory initiatives to large multi-institutional collaborations, rather than the one-size-fits-all approach that was taken with the DataNet program. It is also important that multiple agencies take part in funding projects addressing data management challenges.

A wide array of diverse national-scale initiatives are now providing significant components of an overall research data management solution, but at all scales, data management remains an area of significant dynamism, with new storage technologies, data generation systems, and techniques for data management arising on a regular

[14] See http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504776.

[15] See http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767.

basis. Although work remains to be done in providing national-scale storage infrastructure on which to host the many data collections and data management tools that have been and are being actively developed, the variety of approaches being taken by large-scale projects makes it inevitable that some will be successful and will grow in scope and capability. As experience is gained in meeting data management requirements, infrastructure programs and researchers with data needs will become more sophisticated in their planning and execution of data management.

References

Andrews, Phil, Chris Jordan, and Hermann Lederer. 2006. Design, Implementation, and Production Experiences of a Global Storage Grid. *Proceedings of the 23rd IEEE/14th NASA Goddard Conference of Mass Storage Systems and Technologies*. College Park, MD: IEEE.

Association of American Medical Colleges. 2011. *Challenges and Opportunities for New Collaborative Science Models*. Report from the AAMC Task Force on Information Technology Infrastructure Requirements for Cross-Institutional Research. Washington, DC: AAMC. Available at https://www.aamc.org/download/121174/data/rtfreport-color.pdf.pdf.

Atkins, Daniel E., Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, and Margaret H. Wright. 2003. *Revolutionizing Science and Engineering Through Cyberinfrastructure*. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Available at http://www.nsf.gov/cise/sci/reports/atkins.pdf.

National Science and Technology Council. 2009. *Harnessing the Power of Digital Data for Science and Society*. Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Available at http://www.nitrd.gov/About/Harnessing_Power_Web.pdf.

National Science Foundation. n.d. Dissemination and Sharing of Research Results. Available at http://www.nsf.gov/bfa/dias/policy/dmp.jsp, retrieved August 27, 2013.

Dilemmas of Digital Stewardship: Research Ethics and the Problems of Data Sharing

Lori M. Jahnke and Andrew Asher

Abstract

In a recent qualitative study of data management practices among researchers and faculty members at five universities, we observed that ethical questions are central to scholars' perspectives about the sharing of research data. Ethical questions are especially important for qualitative researchers and for researchers who are working on transnational research teams, which often include individuals and communities with different cultural standards for information sharing. A lack of clear ethical standards and the lack of guidelines for managing privacy and access control in the reuse of research data compound the already complex negotiations involved in managing these diverse constituencies. Universities' increased use of commercial cloud services (e.g., Amazon, Google) as a way to alleviate pressures on staff and budgets for ever greater technical infrastructure and support services further complicates matters by placing yet another piece of the intellectual record under the stewardship of commercial entities. In this article, we explore the ethical dilemmas faced by scholars when sharing their data and the ramifications of outsourcing research infrastructure to the commercial sector. We focus particularly on the effect of these dynamics on data ownership, control, and access, as well as on the degree to which these relationships influence the scholarly process.

Introduction

> We find, then, that there is not a single rule, however plausible, and however firmly grounded in epistemology, that is not violated at some time or other. It becomes evident such violations are not accidental events, they are not results of insufficient

> knowledge or of inattention that might have been avoided. On the contrary, we see that they are necessary for progress (Feyerabend 2010, 7).

The issues of data ownership, control, and access are converging around our increased technical capacity to collect and transmit data, even as we struggle with the social and political aspects of developing the much-needed academic computing infrastructure. As has been demonstrated historically, the emergence of new infrastructure often provokes tensions that erupt in bitter conflicts over decisions that may eventually seem obvious and uncontested. Such conflicts result from the inclusion or exclusion of individuals from resource flows. Over time, cultural change and reorientation of behavior in relation to the emergent infrastructure mask gains and losses within a hybrid of local knowledge and formal structure (Star and Ruhleder 1996, 132).

> Infrastructure, like regulation, may be subject to "capture," in which the interests of powerful established constituencies come to overwhelm and crowd out potential innovations. Infrastructural incumbents may exploit their historically accrued strengths to effectively hold infrastructure in place, stacking the deck against new, less organized, or less favorably placed actors, thereby limiting the scope and vision of new infrastructural possibilities (Edwards et al. 2007, 26).

In other words, disparities of access to resource flows created by the infrastructure become normative over time, and the configuration of the infrastructure is no longer viewed as a source of inertia.

Presumably, the impetus to share research data and build the needed infrastructure emerges from the desire to create a more transparent and collaborative research environment that will allow researchers not only to build directly on earlier work, but also to make more informed decisions regarding future research directions. However, releasing research data is not always straightforward, and building research data infrastructure is fraught with complexities, such as the twin needs to protect sensitive data and to maintain the confidentiality of data on research participants. These issues are particularly important in the social sciences, given their frequent reliance on human participants and qualitative data. Although other research areas may also rely on human participants, the subject of analysis outside the realm of the social sciences is not typically the personal attributes and behaviors of groups or individuals. This distinction in analytical focus presents particular challenges for the task of de-identifying data without destroying their usefulness. Many of the tensions surrounding data sharing are familiar, but the potential scale at which data can be shared via the Internet, combined with a transition to cloud computing, amplifies the ethical concerns of many researchers and presents new threats to sensitive data.

Researchers are expected to maintain the confidentiality of research participants by assessing the risks to the inappropriate release

of sensitive information and guarding against these risks, but they are also ethically bound to publish and release information in a timely manner. In this context, much is left to the discretion of the researcher, which is perhaps appropriate in view of the fact that the researcher is accountable to the funding agency, the public, his or her institution, the profession, and most important, the research participants. In an environment without a data curation infrastructure, the researcher controls the storage and dissemination of research data and, thus, is well positioned to assess the risks of inappropriate data release and its consequences. However, with the trend toward cloud infrastructure for data management, many of the decisions that create risks to privacy have been taken out of the hands of researchers and left to the discretion of institutional or corporate entities. The consequences of these decisions for research participants are as yet unknown. It may be difficult to create accountability within the administrative structure that emerges to support research data curation.

This paper grew out of a qualitative study conducted in 2011–2012 on the data management and curation practices of 23 researchers in the social sciences from five universities[1] (see Jahnke, Asher, and Keralis 2012). The goal of this study was to assess researchers' needs for data management support within their institutions and to gather information on their data practices. Throughout the interviews conducted for this study, researchers expressed a persistent concern about the conflicts between professional ethical standards and compliance with data sharing mandates. The researchers also routinely expressed the need for greater infrastructural support and more access to networked storage for multi-institutional teams. Given the plethora of unresolved security threats related to cloud computing (e.g., Balduzzi et al. 2012; Sood and Enbody 2013), the large-scale adoption of networked storage as a means to facilitate data sharing may be in direct opposition to ethical standards.

Although various articles regarding cloud computing mention ethical concerns, such as the unresolved issue of privacy, these considerations are rarely enumerated. Perhaps there is good reason for this ambiguity, as we have yet to fully appreciate the social implications of the sharing and aggregating of data on a large scale and the effect of the developing infrastructure on confidentiality. In this paper, we examine the ethical conflicts from the researcher's perspective. Drawing on case studies from our previous investigation, we discuss the ways in which digital data curation and data sharing mandates affect the responsibilities of researchers and potentially change their relationship to research participants and the scholarly process. Our discussion of these issues takes place within the context of building data management infrastructure that supports the active data collection phase of the research process, a key component for effectively preserving data at scale. We also discuss how the trend toward outsourcing the infrastructure needed to support digital data

[1] This study was funded by the Alfred P. Sloan Foundation and managed by the Council on Library and Information Resources.

curation may have consequences for institutional expenditures, as well as for the scholarly process.

Vulnerability at Scale: Can We, Should We Manage Research Data Without the Cloud?

In recent years, there has been considerable pressure on university administrators to provide services and infrastructure that will support researchers as they integrate new technologies into their research protocols. The changing support needs come at a time of great economic pressure, which is intensifying the desire of administrators to find new efficiencies and reduce costs. Given that cloud data centers require only one-tenth the administrative staff per server that traditional data centers in higher education require, and that dynamic provisioning of computational resources also requires one-tenth the hardware of traditional computing environments (Wang 2009), the cost savings of cloud data centers are potentially significant. Therefore, outsourcing data center functions to cloud providers may seem like an opportunity to provide a more flexible infrastructure while cutting costs. However, the implications of moving research data to cloud infrastructure (private or public) are not well mapped, and there are a number of unresolved issues surrounding system security and the legal codes, as well as possible hidden costs to the institution.

Cloud computing encompasses a variety of services and infrastructure models that may include public clouds, private clouds, or some hybrid of the two. With a public cloud, the provider maintains a shared service environment that is accessible to any customer; in contrast, a private cloud offers an organization exclusive use of an isolated cloud environment. Although a private cloud may alleviate some of the security concerns associated with a public cloud, the more limited sharing of services associated with a private cloud does not produce the same cost savings. In a hybrid environment, an organization may choose to balance security and cost savings by using a private cloud primarily and using a public cloud only when additional capacity is needed. Outsourcing of cloud computing diminishes not only the researchers' capacity to monitor and assess privacy risks related to their data, but also the institution's capacity to perform audit functions (e.g., security and expenditures). The loss of accountability and transparency within key university functions could have considerable implications for determining the true cost of data management infrastructure, and it is likely to create new "moral hazards" within the institution (Shieber 2009).

Researchers' frustration with the poor usability, lack of features, space restrictions, and the deficiency of cross-institutional collaboration tools provided by university software and networks has resulted in the widespread adoption of consumer-grade cloud storage tools for use in data management. Of our interview participants, 39 percent specifically discussed using these third-party applications in their data management activities, and they were divided almost

evenly between the use of Dropbox and the Google Apps suite, including Google Drive (formerly Google Docs) and Google Mail. The resulting risks to privacy interact with researcher ethics and institutional decision making in a way that could ultimately hinder the development of a robust academic cloud infrastructure.

System Security

When it comes to adopting cloud computing, the desire to catch up with current data practices while simultaneously reducing costs seems to have left security and privacy largely as an afterthought (Glott et al. 2011; Jansen and Grance 2011). Numerous unresolved security issues present a serious conflict for scholars, as the maintenance of data security and privacy is central to helping researchers satisfy their professional codes of ethics and meet the confidentiality standards set by their institutional review boards (IRBs), which are unlikely to review third-party end user agreements for privacy risks.[2]

To make cloud computing solutions attractive, providers create efficiencies by sharing physical and administrative infrastructure among multiple customers. However, the hardware used in this infrastructure (e.g., central processing unit [CPU] caches and graphic processing units [GPUs]) was not typically designed to ensure data isolation within a multitenant architecture (Cloud Security Alliance 2010). Thus, a virtualization hypervisor mediates the tenant operating system's access to the physical resources, allowing multiple tenants to utilize the same physical infrastructure (Ibrahim, Hamlyn, and Grundy 2010).

Virtualization is the key to offering scalable resources, but vulnerabilities in the virtual machine images have allowed attackers to perpetrate malware infections (Balduzzi et al. 2012), gain control over the underlying platform, and inject malicious code that permits the insertion of a rootkit[3] layer below the operating system (Ibrahim, Hamlyn, and Grundy 2010; Sood and Enbody 2013; Vaquero, Rodero-Merino, and Morán 2010). Security researchers have also identified long lists of security issues related to poor user interface security[4] and malicious administrative insiders (for an overview, see Jansen and Grance 2011).

Many of the security concerns surrounding cloud computing systems are not unique to commercial cloud providers (e.g., Amazon, Microsoft,[5] Rackspace's Mozzo), but—more troubling—may be

---

[2] For example, Google Apps for Education, which includes data stored on a university Google Drive account, allows confidential data to be shared with "affiliates" (see http://www.google.com/apps/intl/en/terms/education_terms.html).

[3] A rootkit is a type of software used to enable privileged access to a computer while concealing its existence from normal means of detection. Removal is complicated at best and may even be impossible if it resides in the kernel. A firmware rootkit may require replacement of the affected hardware.

[4] In the analysis of cloud management interfaces by Somorovsky and colleagues (2011), commercial providers responded quickly and attentively to identified issues when possible, but it would be little consolation for researchers to learn that security measures have been taken after their sensitive data may have been exposed.

[5] On October 19, 2012, Microsoft announced that several universities have signed up for their Office365 cloud services.
See http://www.insidehighered.com/news/2012/10/22/universities-and-microsoft-write-standard-privacy-agreement-cloud-services.

endemic to the way that cloud networks are currently constructed. For example, in 2011, Dropbox reported that all files of its users were accessible without a password for about four hours. Dropbox encrypts and decrypts user data on its servers rather than on the user's local computer, allowing multiple devices to be easily synchronized and data to be recovered even if the user loses the password. However, this arrangement also creates vulnerability in that Dropbox can potentially view data and provide them to third parties when required by to do so by legal action. According to Dropbox, less than 1 percent of its (then 25 million) accounts were potentially compromised based on the log-in data from this four-hour period (although users have no way to verify this claim).[6] Such security lapses suggest that researchers should carefully evaluate whether such a service is appropriate for use with their data and whether they should take additional steps, such as local encryption, to protect their sensitive materials. Unfortunately, these additional steps often counteract the ease of use that led researchers to adopt such a service in the first place.

The vulnerabilities of cloud infrastructure can allow the lax security policies of other tenants, data leakage, and service disruptions to propagate through the system and affect multiple tenants (Ibrahim, Hamlyn, and Grundy 2010). Such security lapses are not readily detectable by individual tenants, and necessary functions, such as data isolation within the virtualized environment have yet to be achieved (Subashini and Kavitha 2011). Thus far, commercial providers have offered only limited audit capabilities for their customers, and security incidents are largely invisible to the customer. Problems such as data corruption may not be detected for a long time, and data leakage by skilled insiders is unlikely to be detected at all (Glott et al. 2011). Therefore, an important practical challenge within the cloud computing environment is to develop the abilities (1) to identify and moderate undesired information flows without overstepping the administrative boundaries necessary to maintain confidentiality of the data and (2) to limit administrator misbehavior.

Jurisdiction and Uncertainty

Current legal systems are insufficiently prepared for the challenges that come with the complexity and pervasiveness of cloud computing (Sotto, Treacy, and McLellan 2010). This is a matter of particular concern for researchers in the social sciences, as they are often studying topics that are sensitive to political and social changes that empower or disempower various groups. An inability to control who has access to sensitive data could have severe consequences for research participants who may be members of marginalized groups (e.g., oppressed ethnic groups, victims of abuse, or the mentally ill) or who participate in social movements that are unpopular with the dominant interests. For example, a sociologist we interviewed published on a university blogging site a political op-ed piece discussing

[6] See https://blog.dropbox.com/?p=821.

his experiences in Ethiopia; hackers attempted to sabotage and use the blog against an Ethiopian national living in the United Kingdom. Although the blog did not directly discuss the sociologist's research, had the research data been stored on a cloud server, they might also have come under attack, which, if successful, could have potentially endangered the lives of the research participants.

Unlike local data centers, which are located in a single country, cloud infrastructures often extend over multiple jurisdictions, and it may be unclear which laws apply to the search and seizure of data (Hon, Hörnle, and Millard 2012; Sotto, Treacy, and McLellan 2010). In fact, it is often very difficult, if not impossible, for researchers to determine precisely where their data are held and which laws may apply—an especially problematic situation for researchers who are working transnationally on potentially sensitive topics. To gain access to data that are stored in the United States (e.g., on a hard drive in a research laboratory), law enforcement must present a warrant to search the physical premises. When data are stored in a cloud infrastructure, a subpoena or e-discovery writ may be served to the cloud provider, and the owner of the data may not even be notified of the search. In the United States, Twitter's successful resistance to disclosing private user data is among only a few examples of a company challenging the misuse of secrecy provisions contained in U.S. national security letters (NSL) on behalf of their users.[7] Recently, the *New York Times* reported that the Federal Bureau of Investigation requests more than 50,000 NSLs per year (Cohen 2011). Google has also begun releasing information regarding government requests for user account data and their rate of compliance with these requests (figure 1).

Provider and user needs are not likely to align in relation to disclosure of user activity and resistance to overly broad and unreasonable searches without significant market pressure. Google's Transparency Report is a step in the right direction, but the data released are still very general; they do not allow individual users to assess their risk of data exposure, nor can users determine if their personal data have been released.

It is conceivable that governments might target particular data sets or types of data stored on cloud services without the researcher's knowledge. For example, one anthropologist we interviewed was studying the prosecution of members of the Kurdish minority in Turkey under the auspices of Turkey's anti-terrorism laws. This researcher was very concerned about the possibility of state intrusion into her data, and she was not planning to store them with a cloud data services provider. It is not difficult to imagine a scenario in which government actors compromise a cloud data services provider. It may be advantageous for a data service provider to cooperate

[7] A national security letter (NSL) is a demand letter to communications providers, financial institutions, and credit bureaus to provide certain types of customer business records, including subscriber and transactional information related to Internet and telephone usage, credit reports, and financial records (Yeh and Doyle 2006, 10). An NSL does not need prior approval by a judge, and it may include a gag order to prevent disclosure of the search or even the existence of the order.

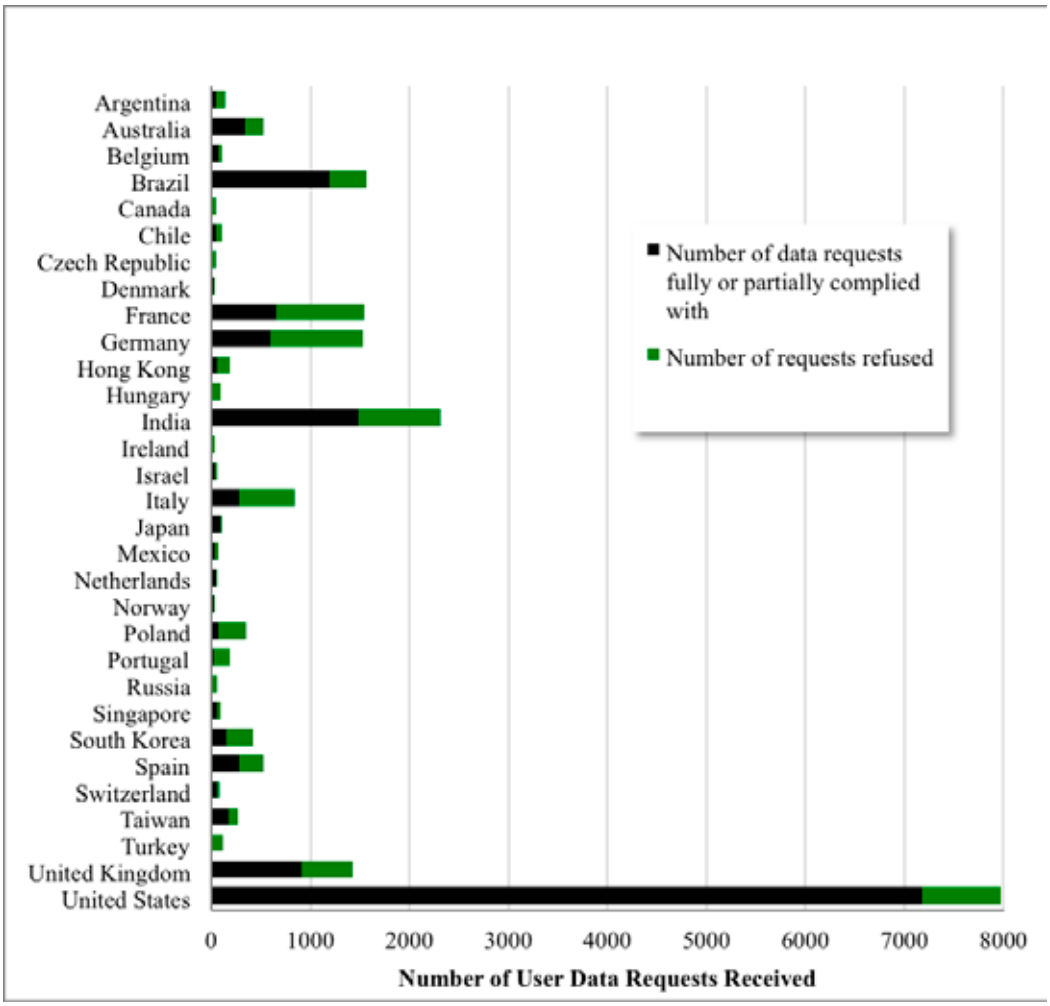

Fig. 1. Requests for Google user data by country from January 2012 to June 2012. Data are from the Google Transparency Report.

with law enforcement rather than risk large-scale service interruptions or seizure of equipment. Moreover, it is not their privacy rights that are in jeopardy. According to Molnar and Schechter (2010), users of cloud infrastructure may be subject to several additional jurisdictional threats (e.g., jurisdictional collateral damage, indirect jurisdictional exposure, and direct jurisdictional exposure), and they advocate the deployment of software tools that enable tenants to manage the hosting location of their applications and data. However, the deployment of such tools would undermine some of the efficiencies that the users were hoping to gain from cloud infrastructure.

Realigning regulations with cloud computing infrastructure would help significantly in providing sound infrastructure for research data. Currently, users of cloud computing services must comply with multiple regulatory provisions simultaneously, and case-by-case agreements with tenants are unlikely to result in a systematic resolution to the problem. The realignment of regulations might also provide additional economies of scale that could benefit both institutions and researchers, such as tracking jurisdictional threats and

disseminating information to help users evaluate jurisdictional risks (Molnar and Schechter 2010). Cloud providers are well positioned to collect and distribute information about jurisdictional threats and changes in legislation, an activity that is costly when performed by individual tenants.

Hidden Costs: Sustainability of Outsourcing Key Research Support Functions

Opting for near-term flexibility and cost reduction over long-term sustainability and cost control may leave universities open to a variety of vendor lock-in and security problems that will ultimately affect budgets and constrain researchers. Although Glott and colleagues (2011) suggest that cloud computing in the future is likely to resemble the seamless federation of Internet service providers that exists today, the cloud computing market currently consists primarily of isolated providers that are motivated to secure market share before their businesses become commodities. (For example, Dropbox's "Great Space Race" was targeted specifically at gaining market share among university-affiliated users.[8]) This market dynamic is not surprising, as cloud computing is still in the early stages of development as infrastructure, despite its similarity to the 1960s conceptualization of the Internet itself (e.g., time-sharing of large computing resources and grid computing that serves large communities).

When compared with that of other large-scale infrastructures in the United States, the growth of the Internet mirrors the pattern of development and the adoption timescale (40–50 years) of other large technical systems (Edwards et al. 2007, 19). The Internet bears many of the hallmarks of the consolidation phase (although the market-oriented nature of access creates widening disparities), but the "virtual infrastructure"[9] that could support research data curation is in a much earlier stage of development. Much of the activity surrounding research data management is characteristic of system building with some experimentation in technology transfer and growth. The legal and social aspects of data curation are also immature with respect to emerging technical capacity. Without judicious decision-making regarding the functions that can be safely outsourced to the commercial sector, university administrators may create costly vendor lock-in issues and delay or circumvent the much needed consolidation phase of infrastructure development for research data.

Because of high switching costs, inferior technologies can become so dominant that even superior technologies cannot surpass them in the marketplace (e.g., the success of the QWERTY keyboard), a phenomenon known as negative path dependence. Ultimately, negative path dependence increases costs to the institution as administrators are forced to consider the costs of staff retraining and

[8] See https://www.dropbox.com/spacerace.

[9] Services built upon existing infrastructures make up the virtual infrastructure. The World Wide Web and e-mail are two examples of virtual infrastructure that rely on the Internet.

reorganization when faced with a provider that is not competitive in either performance or pricing.

> Switching costs can be reduced by making cloud infrastructure more homogenous, such as by using the same cloud hosting APIs and tools as other providers. Alas, such a solution is unlikely as cloud providers use their infrastructures to differentiate themselves from competitors and keep their businesses from becoming commodities (Molnar and Schechter 2010, 8).

The lessons that libraries have learned from their experiences with the evolution of vendor-supported library systems and electronic journals may be quite relevant to this issue, especially given the difficulties encountered in libraries' attempts to transition to open access publication. Thus, unless pressure from stakeholders is significant, we are less optimistic than Glott and colleagues (2011) regarding the evolution of the commercial cloud market into a seamless and robust academic utility.

In addition to the hazards of vendor lock-in, security researchers have already identified weaknesses in the system that may allow cost overrun attacks and deceptive billing practices (either advertent or inadvertent) that may be difficult to track with the limited forensic capabilities currently available (Molnar and Schechter 2010). To address such needs, Glott and colleagues (2011) anticipate that the third-party security market will grow to $1.5 billion by the year 2015 and that it will consume as much as 5 percent of overall IT security spending. These third-party security services represent a future cost of maintaining cloud computing infrastructure that institutions will need to bear. Furthermore, with a continued trend toward outsourcing, the legal costs of managing contracts and policing privacy infractions will also continue to grow.

A related area of expenditure for universities is the on-campus support that must accompany data management systems using cloud infrastructure. Given the numerous security issues, many of which result directly from user behavior, universities will still need to invest in personnel who can help users improve their data practices and can educate users in the proper security measures. The ambiguities surrounding jurisdictional control and the diminished security auditing capacity of the institution make this education and outreach even more important. The opportunity to realign staff with the support needs of university researchers can be seen as a significant benefit to universities, although providing these services is likely to offset much of the cost savings gained by outsourcing cloud infrastructure.

The hidden costs that have been outlined (e.g., vendor lock-in/resistance to system interoperability, post hoc security measures) stem from a mismatch between the mission of higher education institutions and that of commercial cloud providers. The mission of most universities is generally to support scholarship while teaching, preserving, and applying knowledge in the service of humanity. In contrast, commercial providers are motivated to garner the greatest

market share and, therefore, the most profit by the most efficient means possible. There may be many instances in which best practices for teaching, creating, and preserving knowledge do not align with commercial goals. For example, as Molnar and Schechter (2010) have noted, cloud hosting providers are more likely at first to work toward increasing their market share by reducing self-hosting rather than to expend resources to improve the security of the system.

Ethical Dilemmas of Sharing Data

Researchers are facing an increasingly complex milieu of policies governing research practices and the management of the resulting data. Depending on the project, researchers may find themselves dealing with the policies of IRBs, the ethical codes of their professional societies, data sharing mandates of funding agencies, foreign policies governing cultural information, and the intricacies of differing cultural perspectives regarding privacy and the ownership of data. At times, these policies are in direct opposition to one another. It is not surprising, therefore, that researchers expressed uncertainty about the best procedures for adopting data management protocols and making decisions regarding appropriate data access (see Asher and Jahnke 2013). Ethical concerns generally fall into one or more of the following three areas:

1. Protecting the privacy of research participants and maintaining the confidentiality of sensitive information
2. Ambiguity regarding data ownership, particularly in transnational contexts
3. Monitoring and preventing inappropriate use of research data

The researchers we interviewed seemed mostly unaware of the security risks associated with cloud storage, and if they are like most software users, they probably have not read the privacy policies for the tools they use in their research.

Maintaining Privacy of Sensitive Information

Professional codes of ethics vary among the social sciences in their specificity and formality, but they share certain philosophical ideals on the treatment of information related to research participants, as well as on data sharing and the dissemination of research findings. Stated simply, researchers and their teams are expected to treat information provided by research participants as confidential even if there is no legal requirement to do so. They are further expected to use extreme care in transmitting research data and to avoid situations in which there are risks of unauthorized access. For example, the American Sociological Association Code of Ethics states, "Sociologists are attentive to the problems of maintaining confidentiality and control over sensitive material and data when use of technological innovations, such as public computer networks, may open their professional and scientific communication to unauthorized persons" (American Sociological Association 2008, pt. 11.05). The reference

Data Likely to Be Recorded in Social Sciences Research	Data Not Likely to Be Recorded
Names	Social Security numbers
Geographic designators, including geocodes	Telephone and fax numbers
E-mail addresses	Biometric identifiers
URLs and IP Numbers	Medical records numbers
Full-face photographic images	Health plan beneficiary numbers
All elements of dates	Account numbers
All ages over 89	Vehicle identifiers and serial numbers
Any other unique identifying number, characteristics, or codes	Certificate/license numbers
Voice recordings	

Table 1. Data types typically considered identifying information. This table draws on the Emory University IRB Guidelines (Emory University IRB 2012, 293), which includes identifiers that are fairly standard among IRBs.

to computer networks is unusual among professional ethics codes, which seldom directly address data sharing methods and the varying risks associated with different procedures for data storage and transmission (e.g., cloud vs. local storage, private vs. commercial clouds, encryption methods).

If the purpose of an ethics code is to guide decision making when values are in conflict, then researchers are receiving very little support from their professional associations in determining a reasonable course of action for when and how to share their data. The most specific guidance is likely to come from the IRB for those research projects that require its approval (e.g., projects that involve human participants). IRBs typically have specific guidelines for what is considered identifying information (Table 1), as well as for how this information should be protected and managed. The *Institutional Review Board Guidebook*, Chapter V, states:

> If identifiers are recorded, they should be separated, if possible, from data and stored securely, with linkage restored only when necessary to conduct the research. No lists should be retained identifying those who elected not to participate. Participants must be given a fair, clear explanation of how information about them will be handled.
>
> As a general principle, information is not to be disclosed without the subject's consent. The protocol must clearly state who is entitled to see records with identifiers, both within and outside the project. This statement must take account of the possibility of review of records by the funding agency. (OPRR Reports, Dear Colleague Letter (December 26, 1984), p.3, quoted in the IRB Guidebook).[10]

Given the kinds of data typically of interest in the social sciences, data sharing mandates may conflict with these guidelines.

[10] See http://www.hhs.gov/ohrp/archive/irb/irb_guidebook.htm.

When data are curated using a cloud provider, it may become nearly impossible for researchers to determine exactly how their data are stored, to maintain control over access to their data, and to communicate risks of exposure to research participants.

The IRB umbrella has covered social sciences research since the 1960s when Title 45 was enacted.[11] Scholars and policymakers have criticized its authority, however, largely because the regulatory code does not fit the epistemologies of social science disciplines. The code developed as a response to public outrage over abuses of research participants in medical experiments (Schrag 2010).[12] Thus, IRB policies typically assume a Western biomedical context for research. Within this context, social sciences researchers frequently encounter problems revolving around the notion that interviews may cause psychological harm. This idea runs contrary to the findings of numerous studies concluding that participants rarely perceive the discussion of traumatic past experiences as harmful and may, in fact, find the experience therapeutic (e.g., Dyregrov 2004; Griffin et al. 2003).

In reality, the potential for harm to research participants almost always rests with the researcher's inability to protect their privacy and ensure confidentiality of sensitive information. It is on these grounds that many researchers in the social sciences object to the purview of the IRB in their field of research, since consent forms create documentation that cannot be easily protected with respect to privacy. Aside from often being the weakest link in maintaining confidentiality, consent forms can make it difficult to establish trust with research participants, particularly when these individuals have been victims of oppression, may be illiterate, or are otherwise disenfranchised. Although scholars may share the ideals of protecting participant privacy, they may not agree on whose definition of privacy and consent should apply.

The American Sociological Association Code of Ethics mentioned earlier also emphasizes the importance of data sharing as good professional practice, but with the caveat: "They [sociologists] maintain the confidentiality of data, whether legally required or not; remove personal identifiers before data are shared; and, if necessary, use other disclosure avoidance techniques" (American Sociological Association 2008, pt. 13.05c). Given the unresolved problems with security in the cloud, researchers may wish to avoid using data management systems that cannot offer security and auditing guarantees or are inadequately transparent with respect to the risks of data exposure. However, university infrastructure decisions or expediency in meeting research goals necessary for continued employment or tenure may compel them to use such systems.

---

[11] Title 45 (Public Welfare), Part 46 (Protection of Human Subjects) of the Code of Federal Regulations gives authority to IRBs.

[12] Some of the most often cited abuses include Nazi doctors' experiments on Holocaust victims (Shuster 1997); the Tuskegee Syphilis Study, which took place from 1932 to 1972 (http://www.cdc.gov/tuskegee/timeline.htm); the Thalidomide tragedy in Europe during the 1950s (Annas and Elias 1999); and several other experiments discussed by Henry Beecher (1966).

An environmental studies scholar who uses face-to-face interviews and secondary data sets to collect quantitative and qualitative data in Kyrgyzstan exemplifies some of these privacy concerns (see Jahnke, Asher, and Keralis 2012). Although funding agencies did not require this researcher to have a data sharing or data management plan, she would like to preserve her data and make it public for the benefit of policymakers and other scholars. However, the data cannot be released in its raw form because it includes voice recordings of interviews and other potentially identifying information. The interviews could be transcribed and then scrutinized for identifying information prior to release, but the need to transcribe from three languages (Kyrgyz, Russian, and English) has made it difficult to find qualified transcriptionists. As a result, it has taken several years to complete only the transcription of the data, and review of the data for identifying information remains incomplete. Given the amount of time and labor that must be invested before these data can be released appropriately, important contextual information may be lost before their release. This scholar needs the support of a data management environment that can handle multiple access levels for an integrated data set, as well as enable fine-grained assessment and management of confidentiality risks.

Determining Data Ownership

The problems surrounding data ownership are nuanced, particularly when conducting research with object collections in a transnational context. Museums and national cultural institutes may have very different perspectives from one another and the researcher on who can own and release images or other raw data related to their collections, and access to collections is often contingent on guidelines specified for each project. A biological anthropologist in our study outlined the difficulties of potentially making high-resolution computerized tomography (CT) scans of bones available through a website. To find an appropriate comparative collection of chimpanzee bones, the researcher had to use a museum collection from Belgium, which complicated the release of the data set. The museum may consider the bone scans proprietary and assert ownership over data produced from their collections and with their equipment. In this case, data rights could become a source of conflict, as the researcher's institution asserts ownership over data produced by university-owned scanners, which were used to create another portion of the data set. Ambiguity surrounding such conflicts can make it difficult to curate data sets and can cause researchers to be very conservative in their data sharing practices.

Perspectives regarding data ownership may evolve over time as object repositories become more accustomed to the need for digital data curation and sharing. For now, however, there is little resolution in this area. As repositories increase their capacity for digital representations of objects, some of the issues surrounding ownership of proprietary data sets may be resolved, or at least clarified.

Monitoring Inappropriate Use of Research Data

Even after appropriate measures have been taken to de-identify data, unmonitored access to a data set could allow it to be aggregated and cross-referenced with additional data that might make it possible to re-identify participants. For example, in the United States, 87 percent of adults can be positively identified with only three data points: their five-digit postal code, their birth date, and their sex (Sweeney 1997). Using the Cambridge, Massachusetts, voter rolls, Sweeney was able to identify 97 percent of individuals using only two data points: their birth date and their full postal code. These tests were based on demographic data with which she demonstrated that knowledge brought to the data by the recipient could be used to reconstruct identifying characteristics, a phenomenon later termed the "power of the adversary" (El Emam et al. 2012, 11). In a large and diverse database, attributes of various fields, such as dominant age classes or ethnicities within a geographic area, can be used to create subsets of anomalous data (i.e., small groups of individuals who do not adhere to the demographic norm for the area). Although the data may appear anonymous by virtue of being part of a large and diverse data set, individuals within the anomalous groups are more likely to be identified because of their rarity within a subset.

By the time Sweeney published her study in 1997, numerous abuses of personal information had already been documented. In a survey of Fortune 500 corporations conducted by Linowes and Spencer (1989), 35 percent of respondents had used medical records to make decisions about employees (n = 3.7 million employees at 126 companies). In another example recounted by Woodward (1995), a banker cross-referenced a list of cancer patients against a list of those who had outstanding loans at his bank; he then called in the loans for individuals appearing on both lists.

Fifteen years after Sweeney's 1997 article, volumes of personal, demographic, and geographic information flood the Internet, and the average individual has access to significantly more computing power than in the mid-1990s (Hilbert and López 2011)—an amount that is likely dwarfed by the concentration of resources available to corporate and government entities for reprocessing data. Even though standards for the de-identification of personal information have become more stringent (e.g., those related to health information), the overall success rate for re-identification is still high, 34 percent for health data and 26 percent for other types of data (El Emam et al. 2011; see also Narayanan, Gong, and Song 1995). In addition, researchers in our study who work with object collections expressed reluctance to release data that could be used to locate rare and valuable items in museum collections, archives, and archaeological sites for fear of exposing them to theft.

Conclusion: Implications for the Scholarly Process

The cornerstone of ethical responsibility to research participants is that of privacy, but this responsibility and the present requirements of sharing data in a digital environment are almost mutually exclusive. As Somorovsky and colleagues demonstrated, the complexity of cloud computing and the ease with which weaknesses can be exploited create a "large seedbed of potential vulnerabilities" (2011,11). The authors concluded that control interfaces are likely to be an attractive target for organized crime in the near future, which could have severe consequences for researchers and their data. Even researchers who do not collect sensitive data could become collateral damage in attacks directed at data stored on the same server. Once researchers understand the extent and the intractable nature of some of the security vulnerabilities, as well as the threat to the privacy of research participants, they may begin to refuse data sharing as a matter of conscientiousness.

> Short-term experiences of gain and loss will shape the incentive structures of individuals and institutions tasked with responding to infrastructural change. This in turn will shape the climate within which infrastructures struggle to emerge: broadly receptive, with allies adding support and innovation to extend the reach, quality, and fit of infrastructure? Or openly or covertly hostile, with important user groups and audiences dragging their heels, undermining change, putting forward counter-projects, or simply refusing to play along? Failing to think proactively about the distributional consequences of infrastructure is not only bad politics, but bad business (Edwards et al. 2007, 24–25).

Given the difficulties of maintaining privacy in a digital environment, we must ask whether research participant privacy is possible in an environment of shared, aggregated data. Continued neglect of the unresolved security and privacy threats in cloud computing will move us even farther into a system in which there is no privacy—a move that will compel us to rethink ethical norms or to abandon the notion of privacy altogether.

Reconfiguring ethical norms away from an emphasis on research participant privacy would surely represent a loss for researchers and the public. If a researcher cannot safeguard the privacy of participants or accurately communicate the risks of exposure, individuals may no longer be willing to become research subjects. For participants, the desire to protect their privacy may quickly overwhelm the desire to contribute to the public good. Inattention to the gains and losses of researchers as users of the cloud could have significant consequences for innovation within the scholarly process. The social costs of disenfranchising researchers from the means to satisfy their ethical responsibilities are as yet unknown.

Despite the difficulties in navigating the social and technical spaces in which the issues surrounding data curation reside, the growth of a robust academic computing infrastructure is likely to

enable new possibilities for researchers. However, in order to create a system that can efficiently manage research data, we will need to make difficult decisions regarding acceptable levels of deviation from standards regarding privacy and otherwise. If care is not taken to accommodate flexibility and resilience to new discoveries within research data infrastructure, we are at risk of fixing research methodologies in time and creating a systematic reinforcement of conservatism while also undermining fundamental responsibilities to research participants. Will it be possible, in this context, to foster the intellectual freedom that has been so important for scientific discovery, or will we inadvertently reward conformity of method and thought? The once revolutionary ideas that transform into obstacles of thought[13] may be buttressed not only by the professional prestige to which we are accustomed, but also by an additional technical and social bureaucracy. As Feyerabend said, "Variety of opinion is necessary for objective knowledge. And a method that encourages variety is also the only method that is compatible with a humanitarian outlook" (Feyerabend 2010, 25).

Recommendations for University Administrators

In order to better address the ethical dilemmas faced by social science researchers when managing, preserving, and making publically accessible their data sets, we make the following recommendations to university administrators:

- Data curation systems should be adapted to the actual behaviors of researchers rather than to idealized or stereotyped research behaviors. Planning for behavioral idiosyncrasy may also point the way to a more robust type of user interface security.
- Neglecting privacy concerns may make researchers even more reluctant to share data. Thus, university administrators must work diligently to create transparency in developing data curation infrastructure and must collaborate with researchers to ensure that privacy requirements are met.
- University administrators should be skeptical of the apparent lower costs of outsourcing cloud infrastructure in light of prior experiences with costly vendor lock-in problems. Additionally, to achieve a satisfactory level of security, institutions will need to make significant investments in staff retraining and researcher education, as well as to implement auditing protocols to ensure the health of the system.
- Institutional representatives should collaborate cross-institutionally to address the lagging legal codes that leave researchers and their data exposed to a variety of jurisdictional threats. Through advocacy for updated privacy legislation, universities have an opportunity to contribute substantially to the public good while supporting their researchers.

[13] Several historians of science have discussed this phenomenon (see Feyerabend 2010; Khun 2012).

References

American Sociological Association. 2008. *American Sociological Association Code of Ethics and Policies and Procedures of the ASA Committee on Professional Ethics.* Washington, D.C. doi:10.1111/j.0028-1425.2007.ethics.x.

Annas, George J., and Sherman Elias. 1999. Health Law and Ethics Thalidomide and the Titanic: Reconstructing the Technology Tragedies of the Twentieth Century. *American Journal of Public Health* 89 (1): 98–101.

Asher, Andrew, and Lori M. Jahnke. 2013. Curating the Ethnographic Moment. *Archive* 3 (Summer). Available at http://www.archivejournal.net/issue/3/archives-remixed/curating-the-ethnographic-moment/

Balduzzi, Marco, Jonas Zaddach, Davide Balzarotti, and Sergio Loureiro. 2012. A Security Analysis of Amazon's Elastic Compute Cloud Service. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 1427–1434. New York: Association for Computing Machinery.

Beecher, H. K. 1966. Ethics and Clinical Research. *New England Journal of Medicine* 274 (24): 1354–1360.

Cloud Security Alliance. 2010. *Top Threats to Cloud Computing* V1.0. Available at https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf.

Cohen, Noam. 2011. Twitter Shines a Spotlight on Secret F.B.I. Subpoenas. New York Times, January 9. Available at http://www.nytimes.com/2011/01/10/business/media/10link.html?_r=0.

Dyregrov, Kari. 2004. Bereaved Parents' Experience of Research Participation. *Social Science & Medicine* 58 (2):391–400. doi:10.1016/S0277-9536(03)00205-3.

Edwards, Paul N., Steven J. Jackson, Geoffrey C. Bowker, and Cory P. Knobel. 2007. Understanding Infrastructure: Dynamics, Tensions, and Design. Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures." Ann Arbor: University of Michigan, School of Information.

El Emam, Khaled, Luk Arbuckle, Gunes Koru, Benjamin Eze, Lisa Gaudette, Emilio Neri, Sean Rose, Jeremy Howard, and Jonathan Gluck. 2012. De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset. *Journal of Medical Internet Research* 14 (1):e33. doi:10.2196/jmir. 2001.

El Emam, Khaled, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. 2011. A Systematic Review of Re-identification Attacks on Health Data. *PloS One* 6 (12): e28071. doi:10.1371/journal.pone.0028071.

Emory University IRB. 2012. *Emory University Institutional Review Board Policies and Procedures*. Revised August 16, 2012.

Feyerabend, Paul. 2010. *Against Method*. 4th ed. Brooklyn, NY: Verso.

Glott, R., Elmar Husmann, A. R. Sadeghi, and M. Schunter. 2011. Trustworthy Clouds Underpinning the Future Internet. *Future Internet Assembly* LNCS 6656:209–221. doi:10.1007/978-3-642-20898-0_15.

Griffin, Michael G., Patricia A. Resick, Angela E. Waldrop, and Mindy B. Mechanic. 2003. Participation in Trauma Research: Is There Evidence of Harm? *Journal of Traumatic Stress* 16 (3): 221–227. doi:10.1023/A:1023735821900.

Hilbert, Martin, and Priscila López. 2011. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science* 332 (6025): 60–65. doi:10.1126/science.1200970.

Hon, W. K., Julia Hörnle, and Christopher Millard. 2012. Data Protection Jurisdiction and Cloud Computing–When Are Cloud Users and Providers Subject to EU Data Protection Law? The Cloud of Unknowing. *International Review of Law, Computers and Technology* 26 (2-3): 129–164.

Ibrahim, Amani S., James Hamlyn, and John Grundy. 2010. Emerging Security Challenges of Cloud Virtual Infrastructure. In *Proceedings of APSEC 2010 Cloud Workshop*. Sydney, Australia.

Jahnke, Lori, Andrew Asher, and Spencer D. C. Keralis. 2012. *The Problem of Data*. Washington, D.C.: Council on Library and Information Resources.

Jansen, Wayne, and Timothy Grance. 2011. *Guidelines on Security and Privacy in Public Cloud Computing* (Special Publication 800-144). Gaithersburg, MD: National Institute of Standards and Technology. Available at http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf.

Khun, Tomas S. 2012. *The Structure of Scientific Revolutions*. 4th ed. Chicago: University of Chicago Press.

Linowes, D. F., and R. C. Spencer. 1989. Privacy: The Workplace Issue of The 90's. *John Marshall Law Review* 23: 591–620.

Molnar, David, and Stuart Schechter. 2010. Self Hosting vs. Cloud Hosting: Accounting for the Security Impact of Hosting in the Cloud. In *Proceedings of the Ninth Workshop on the Economics of Information Security (WEIS)*, 1–18. Redmond, Wash.: Microsoft Research.

Narayanan, Arvind, Neil Zhenqiang Gong, and Dawn Song. 1995. On the Feasibility of Internet-Scale Author Identification. *2012 IEEE Symposium on Security and Privacy* (SP): 300–314.

Schrag, Zachary M. 2010. *Ethical Imperialism: Institutional Review Boards and the Social Sciences*, 1965–2009. Vol. 3. Baltimore, Md.: Johns Hopkins University Press.

Shieber, Stuart M. 2009. Equity for Open-access Journal Publishing. *PLoS Biology* 7 (8): e1000165. doi:10.1371/journal.pbio.1000165.

Shuster, Evelyne. 1997. Fifty Years Later: The Significance of the Nuremberg Code. *New England Journal of Medicine* 337 (20): 1436–1440.

Somorovsky, Juraj, Mario Heiderich, Meiko Jensen, Jorg Schwenk, Nils Gruschka, and Luigi Lo Iacono. 2011. All Your Clouds Are Belong to Us—Security Analysis of Cloud Management Interfaces. In *Proceedings of the 3rd ACM Workshop on Cloud Computing Security* (pp. 3-14). New York, NY: ACM.

Sood, Aditya K., and Richard J. Enbody. 2013. Targeted Cyber Attacks: A Superset of Advanced Persistent Threats. *IEEE Security and Privacy* 11(1):54-61. doi:10.1109/MSP.2012.90.

Sotto, Lisa J., Bridget C. Treacy, and Melinda L. McLellan. 2010. Privacy and Data Security Risks in Cloud Computing. *Electronic Commerce & Law Report* 15: 186.

Star, Susan Leigh, and Karen Ruhleder. 1996. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research* 7 (1): 111–134.

Subashini, S., and V. Kavitha. 2011. A Survey on Security Issues in Service Delivery Models of Cloud Computing. *Journal of Network and Computer Applications* 34 (1): 1–11. doi:10.1016/j.jnca.2010.07.006.

Sweeney, L. 1997. Weaving Technology and Policy Together to Maintain Confidentiality. *The Journal of Law, Medicine & Ethics* 25 (2-3): 98–110.

Vaquero, Luis M., Luis Rodero-Merino, and Daniel Morán. 2010. Locking the Sky: A Survey on IaaS Cloud Security. *Computing* 91 (1): 93–118. doi:10.1007/s00607-010-0140-x.

Wang, Rex. 2009. Cloud Computing: Separating Hype from Reality. *Oracle Keynote, Cloud Computing Conference and Expo* November 4, 2009. Santa Clara, CA. http://www.slideshare.net/wrecks/oracle-keynote-cloud-expo-110409.

Woodward, Beverly. 1995. The Computer-Based Patient Record and Confidentiality. *New England Journal of Medicine* 333: 1419–1422.

Yeh, Brian T., and Charles Doyle. 2006. USA PATRIOT Improvement and Reauthorization Act of 2005: A Legal Analysis. CRS Report for Congress (RL33332). Library of Congress Washington, D.C. Congressional Research Service.