8-2012

The Problem of Data

Lori Jahnke

Andrew Asher
Bucknell University, andrew.asher@bucknell.edu

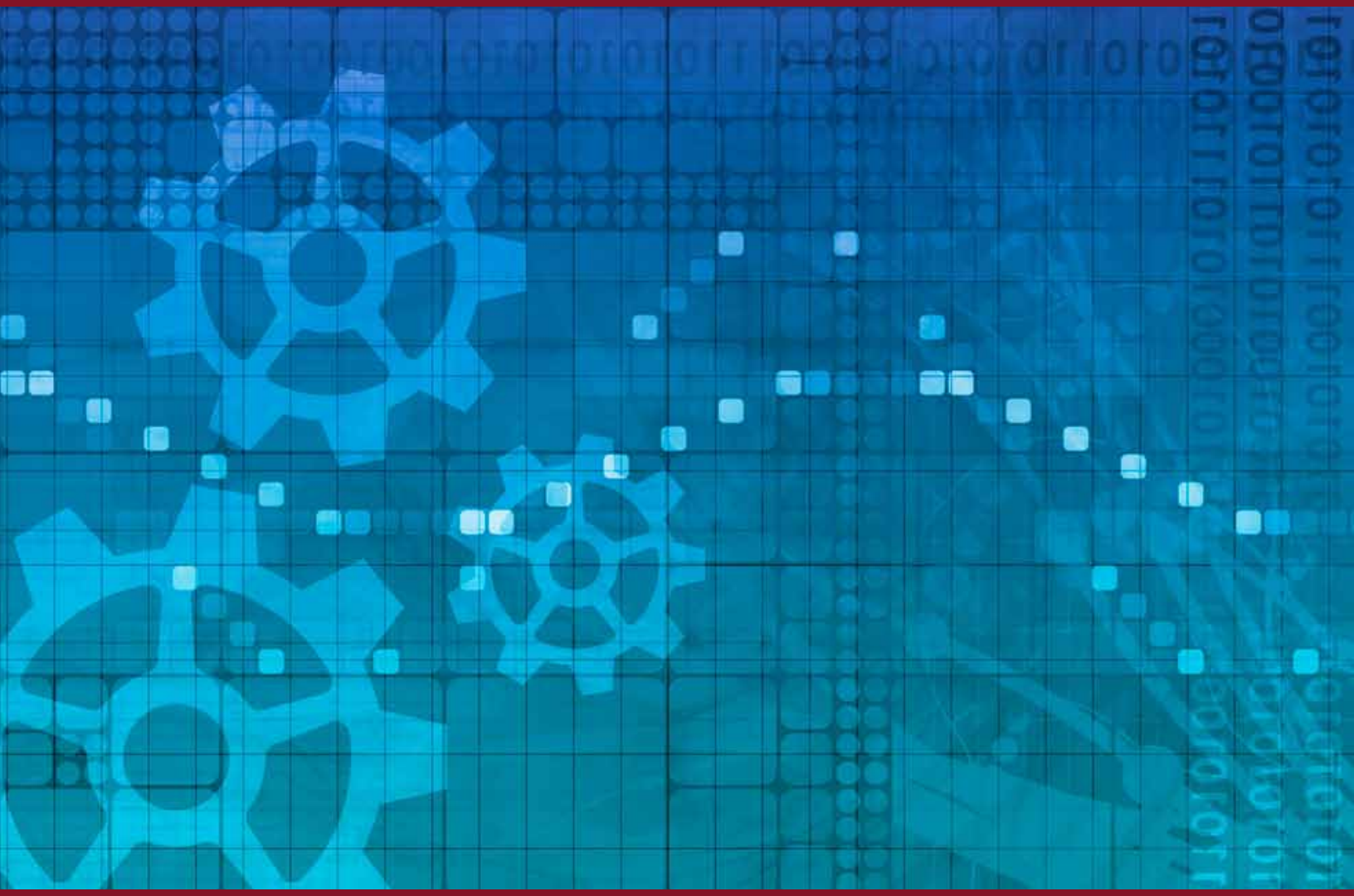
Spencer D.C. Keralis

Recommended Citation
Lori Janke, Andrew Asher, and Spencer D.C. Keralis (2012) The Problem of Data. Council on Library and Information Resources (CLIR) Report, pub. #154. ISBN 978-1-932326-42-0

The Problem of Data

Lori Jahnke and Andrew Asher
Spencer D. C. Keralis
with an Introduction by Charles Henry

August 2012



Council on Library and Information Resources | Digital Library Federation

The Problem of Data

Lori Jahnke and Andrew Asher
Spencer D. C. Keralis
with an introduction by Charles Henry

August 2012

Council on Library and Information Resources

Digital Library Federation
Program of the Council on Library and Information Resources

ISBN 978-1-932326-42-0
CLIR Publication No. 154
Published by:

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

Contents

About the Authors

Andrew Asher is digital initiatives coordinator and scholarly communications officer at Bucknell University, where he leads the university's open access program 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, utilize, and preserve their research data.

Charles Henry is president of the Council on Library and Information Resources. He is also a board member of the National Institute for Technology in Liberal Education (NITLE) and of the Center for Research Libraries, and is a member of the Scientific Board of the Open Access Publishing in the European Network (OAPEN) project. In collaboration with NITLE, he is currently publisher of Anvil Academic Publishing, which focuses on new forms of scholarly research and expression.

Lori Jahnke is Lori Jahnke is currently the anthropology librarian at Emory University. She was previously a CLIR postdoctoral fellow at The College of Physicians of Philadelphia and the University of Pennsylvania. Her primary project was to contribute to the development of the Medical Heritage Library as a multi-institutional collaboration for digitization in the health sciences. In addition to her work in libraries and digitization, Jahnke is a practicing anthropologist.

Spencer D. C. Keralis is director of the Digital Scholarship Co-Operative (DiSCo) at the University of North Texas (UNT). His current research focuses on the implications of social media, digital curation, and data management for the future of the humanities. From 2011 to 2012, he was a CLIR postdoctoral fellow with the UNT Libraries.

Introduction

Charles Henry

> **"Every day, we create 2.5 quintillion bytes of data—so much that 90% of the data in the world today has been created in the last two years alone."**
>
> —IBM, *Bringing Big Data to the Enterprise*[1]

This extraordinary and often cited statistic is an apt quantitative introduction to our technological era, increasingly referred to as the era of Big Data. The massive scale of data creation and accumulation, together with the increasing dependence on data in research and scholarship, are profoundly changing the nature of knowledge discovery, organization, and reuse. As our intellectual heritage moves more deeply into online research and teaching environments, new modes of inquiry emerge; digital data afford investigations across disciplinary boundaries in the sciences, social sciences, and humanities, further muddling traditional boundaries of inquiry.

How then are we responding to what may be the most complex and urgent contemporary challenge for research and scholarship? With considerable difficulty, as the two reports in this volume attest. The key focus of these reports—"The Problem of Data: Data Management and Curation Practices Among University Researchers," by Lori Jahnke and Andrew Asher, and "Data Curation Education: A Snapshot," by Spencer Keralis—is data curation, a term generally defined as a set of activities that includes the preserving, maintaining, archiving, and depositing of data to keep it secure, intact, and accessible for reuse. The term can also comprise the conceptualization and creation of digital objects. In this respect, data curation encompasses the life cycle of data from their inception to their reuse to their transformation into new knowledge products.

---

[1] http://www-01.ibm.com/software/data/bigdata/

Two phenomena compound the challenge of data curation. First, although the stewardship of digital data demands both general and domain specialist knowledge, there are currently no effective ways to prepare people for that hybrid role. Still a developing practice, digital curation has thus far drawn individuals with varied professional experience; many have had no specialist training in the disciplines that they now serve. According to the Digital Curation Centre in Edinburgh, the result is "a shortage of experienced data scientists and curators with digital preservation experience."[2]

The second phenomenon compounding the challenge is the lack of conformity among the places of practice. Libraries, data centers, academic departments—all organizations where data curation can be done—have varied, sometimes idiosyncratic, approaches and often entail different attitudes, cultures, and practices. New government requirements for exposing and managing federally funded research data add urgency to the challenge of curating data.

These two reports address each of these circumstances in depth. Jahnke and Asher explore workflows and methodologies at a variety of academic data curation sites, and Keralis delves into the academic milieu of library and information schools that offer instruction in data curation. Their conclusions, while not surprising, nonetheless point to the urgent need for a reliable and increasingly sophisticated professional cohort to support data-intensive research in our colleges, universities, and research centers. We will need more innovative approaches to recognize, educate, promote, and retain those individuals who evidence the complex skill sets required for the demands of data curation. At the same time, we will need to foster and facilitate a greater coherence of practices, standards, and protocols among the various data sites.

CLIR and the Digital Library Federation have received a major grant from the Alfred P. Sloan Foundation to develop the cohort needed and to help instantiate best practices and shared methods across data curation centers. The grant, made in response to the findings of the reports that follow, could not be more timely. As the recently published report, *One Culture*, asserts, we are now confronted with a new paradigm: a digital ecology of data, algorithms, metadata, analytical and visualization tools, and new forms of scholarly expression.[3] The implications of this digital milieu for the practices of research, teaching, and learning, as well as for the economics and management of higher education, should be of profound interest not only to researchers engaged in computationally intensive work, but also to college and university administrations, scholarly societies, funding agencies, research libraries, students, and academic publishers.

In this respect, we are only just getting started.

---

[2] http://www.dcc.ac.uk/about-us/dcc-charter.

[3] Williford, Christa, and Charles Henry. 2012. *One Culture: Computationally Intensive Research in the Humanities and Social Sciences*. A Report on the Experiences of First Respondents to the Digging Into Data Challenge. Washington, DC: Council on Library and Information Resources. Available at http://www.clir.org/pubs/reports/pub151.

The Problem of Data: Data Management and Curation Practices Among University Researchers

Lori M. Jahnke and Andrew Asher

EXECUTIVE SUMMARY

CLIR was commissioned by the Alfred P. Sloan Foundation to complete a study of data curation practices among scholars at five institutions of higher education. We conducted ethnographic interviews with faculty, postdoctoral fellows, graduate students, and other researchers in a variety of social sciences disciplines. The goals of the study were to identify barriers to data curation, to recognize unmet researcher needs within the university environment, and to gain a holistic understanding of the workflows involved in the creation, management, and preservation of research data.

Key Findings

- None of the researchers interviewed for this study have received formal training in data management practices, nor do they express satisfaction with their level of expertise. Researchers are learning on the job in an ad hoc fashion.
- Few researchers, especially among those who are early in their career, think about long-term preservation of their data.
- The demands of publication output overwhelm long-term considerations of data curation. Metadata and documentation are of interest only if they help a researcher complete his or her work.
- There is a great need for more effective collaboration tools, as well as online spaces that support the volume of data generated and provide appropriate privacy and access controls.
- Few researchers are aware of the data services that the library might be able to provide and seem to regard the library as a

We would like to thank the Alfred P. Sloan Foundation for its generous funding, which enabled us to carry out this study. Additional thanks goes to our many colleagues at CLIR who provided insightful commentary and support.

dispensary of goods (e.g., books, articles) rather than a locus for real-time research/professional support.

Recommendations

- There is unlikely to be a single out-of-the-box solution that can be applied to the problem of data curation. Instead, an approach that emphasizes engagement with researchers and dialog around identifying or building the appropriate tools for a particular project is likely to be the most productive.
- Researchers must have access to adequate networked storage. Universities should consider revising their access policies to support multi-institutional research projects.
- Educational or other training programs should focus on early intervention in the researcher career path for the greatest long-term benefit.
- Data curation systems should be integrated with the active research phase (i.e., as a backup and collaboration solution).
- In the area of privacy and data access control, additional tools should be developed to manage confidential data and provide the necessary security. Most importantly, policies must be developed that support researchers in this use of these technologies.
- Many researchers expressed concerns surrounding the ethical reuse of research data. Additional work is needed to establish best practices in this area, particularly for qualitative data sets.

INTRODUCTION

By 1977 print media had already begun to show signs that its relevance was declining in relation to electronic media (Pool 1983). However, it was only after 2000 that digital storage formats became a significant portion of total storage media, and by 2007, 94 percent of technological memory was in digital format (Hilbert and López 2011). Although digital technologies have brought new opportunities for researchers to create data sets that enable increasingly sophisticated analyses, haphazard data management and preservation strategies endanger the benefits that this advancement might bring. Although digital data curation in its most basic form is merely saving the bits and bytes, the underlying ethical and philosophical issues related to sharing data amplify the technological challenge at hand. It is essential to address these issues in order to develop policies and infrastructure that truly support scholars in this new era.

The tasks associated with conducting research under a data-intensive paradigm increase the pressure on already overextended research schedules. Scholars are also grappling with the ethical and philosophical problems of data sharing in a vacuum of coherent policy support for data linking and release. The purpose of this study is to gather a more complete and researcher-centered understanding of the data usage, management, and preservation practices of university-level faculty, postdoctoral researchers, and staff researchers. Our

goals were to identify barriers to data curation within the university environment, as well as to

- gain a holistic understanding of the workflows involved in the creation, management, and preservation of research data;
- identify unmet researcher needs within these processes; and
- use this information to make curricular, policy, and funding recommendations for data curation practices.

We conducted ethnographic interviews at five institutions with researchers from a variety of disciplines in the social sciences (table 1; see Appendix A for an overview of the data). The interviews focused on how the researchers collect and analyze data; how they manage, preserve, and archive these data; and what training they have had in data curation practices (Appendix B).

BACKGROUND

The rapid shift in the materiality of data has had tremendous consequences for researchers and their products. As Mathews and colleagues note, "Simple notions of access are substantially complicated by shifting boundaries between what is considered information versus material, person versus artifact, and private property versus the public domain" (2011, 725). These researchers are referring to the use of stem cell lines in research and the inherent ambiguity of navigating privacy and consent when the research materials are both human-made and derived from human individuals. Issues of privacy and consent are no less relevant to social scientists. As the lines around research materials continue to blur, so do disciplinary boundaries, thus necessitating careful discussion of data access and security. King observes:

> [P]arts of the biological sciences are effectively becoming social sciences, as genomics, proteomics, metabolomics, and brain imaging produce large numbers of person level variables, and researchers in these fields join in the hunt for measures of behavioral phenotypes. In parallel, computer scientists and physicists are delving into social science data with their new methods and data-collection schemes (King 2011, 719).

The practical applications for integrating data from diverse yet complementary fields are numerous. For example, synthesizing social science, ecological, and hydrological data could help society cope with climate change (Overpeck et al. 2011), design better cities (Gur et al. 2011), and improve public health and the delivery of care. Standardizing and linking data from demographic studies, health surveillance systems, and pathogen-related studies could significantly improve the delivery of health care in remote areas that lack local medical expertise (Lang 2011).

Thoughtfully integrated pools of data could also promote transparency in research (Gur et al. 2011) and improve research methodologies by enabling the identification of unstated assumptions or

theories that shape analytical outcomes (Rzhetsky et al. 2006; Smail 2008). Cokol and colleagues (2005) have demonstrated that in the sciences researchers tend to focus on established areas of knowledge rather than testing novel approaches and methods. As a result, popular fields may be overstudied while other lines of inquiry may be neglected entirely. Evans and Foster (2011) argue that a meta-analysis of research findings (i.e., publications) could identify overstudied fields where continued research has diminishing returns, thus helping individuals make better decisions about research investment. Aggregated research data could make such efficiencies clear. Failed investigations rarely receive the attention of a publication, but they do generate data that may indicate invalid approaches or the lack of merit in a particular line of inquiry. The aversion to publishing less than stellar outcomes leads to a tremendous duplication of scholarly effort.

Despite its advantages, integrating data from multiple fields is not without risk to the preservation of the intellectual rigor of academe. In the field of neuroscience, for example, Akil and colleagues (2011) suggest that integrating neural connectivity data with behavioral phenotype data (e.g., IQ scores) will provide new insight into the spatial organization and function of the human brain. This may be true. However, the validity of intelligence testing is a notoriously contentious topic, and the concept of intelligence is rather subjective (Nisbett 2003). Anthropologists and feminists, among others, have long disputed the validity of IQ tests, as well as the merit of characterizing psychological properties as human universals when the experiments frequently rely on American undergraduates as the research subjects.[1] Henrich and colleagues (2010) have shown that the universality of undergraduate cognition is a false assumption. In fact, they conclude that WEIRD (Western Educated Industrialized Rich and Democratic) subjects are among the least representative populations for characterizing the fundamentals of human psychology. Without sufficient attention to the context of data aggregated from an array of fields, we run the risk of promoting facile interpretations of the relationship between human biology and behavior, and of human nature itself.

The form and quantity of information available could make possible significant advancement in addressing societal problems, if we can provide sustainable infrastructure and formulate the coherent policies needed to support it. The data deluge leaves us with several big questions; the answers will help define individual privacy rights, personhood, electronic identity, and our relationship to these concepts. We face a tremendous challenge in preserving the vast amounts of research data while balancing the need to protect sensitive data with the need to provide meaningful access for researchers and other stakeholders. This task cannot be accomplished without the investment of the researchers themselves.

[1] According to the analysis of Arnett (2008), 67 percent of American studies published during the year 2007 in the *Journal of Personality and Social Psychology* drew their research subjects from the pool of psychology undergraduates. For non-American studies the rate was even higher, 80 percent (Arnett 2008, 604). For a more extensive discussion and critique of this issue, see Henrich et al. (2010).

Site	Type*	Size*	Enrollment Classification*	Research Classification*	Disciplines/Areas of Study
Penn State University	Public	45,185	High undergraduate	Research university/very high activity	Biological Anthropology, Archaeology, Sociology Education, Slavic Languages
Lehigh University	Private, not-for-profit	6,996	High undergraduate	Research university/high activity	Psychology, Education, Political Science, Architectural History
Bucknell University	Private, not-for-profit	3,737	Very high undergraduate	Baccalaureate/arts and sciences	Political Science, Sociology, Environmental Science, International Relations, Anthropology
Johns Hopkins University	Private, not-for-profit	20,383	Majority graduate/ professional	Research university/very high activity	Sociology and Public Policy, Applied Mathematics, Geology (data scientist), Sociology, Anthropology
University of Pennsylvania	Private, not-for-profit	24,599	Majority graduate/ professional	Research university/very high activity	Education, Archaeology, History

Table 1. Characteristics of research sites included in this study

* Information from Carnegie Foundation for the Advancement of Teaching (2010).

RESEARCHER PERSPECTIVES AND UNMET NEEDS

Several studies have acknowledged that rates of coauthorship are increasing across academe and that the collaborative laboratory work model is replacing that of the lone scholar-genius (King 2011). Although there are certainly larger social and economic factors at work here, undoubtedly access to more data is changing not only the way that social scientists work, but also the kinds of questions that they can investigate. The incorporation of data from a variety of sources to address a single research problem causes the proliferation of roles within the research team without clear avenues of support and training.

Participants in this study included researchers in the social sciences of various ranks, but the focus was on early career professionals. We initially planned to interview postdoctoral fellows, junior rank faculty, and researchers exclusively, but it quickly became apparent that the challenges of digital data curation are spread widely throughout the scholarly workforce and that issues related to rank and training are interwoven with the complexities of multidisciplinary research teams. In the hope of providing a broader perspective on unmet needs in academe and changing needs over time, we also included a few advanced graduate students, as well as senior faculty (table 2).

Participant #	Rank/title	Ph.D. Discipline
1-03-100511	Assistant Professor	Anthropology (Biological)
1-17-121211	Assistant Professor	Education
2-12-111011	Assistant Professor	Environmental Science
2-16-120211	Assistant Professor	Anthropology
3-05-102111	Assistant Professor	Developmental Psychology
3-07-102111	Assistant Professor	Political Science
3-08-102111	Assistant professor	Political Science
2-22-021512	Professor	Environmental Studies
2-15-120211	Associate Professor	International Relations
5-09-103111	Data Scientist	Geology
1-01-72911	Digital Curator	Slavic Languages
5-10-103111	Graduate Student	Sociology
5-20-020212	Graduate Student	Environmental engineering
5-21-02032011	Graduate Student	Anthropology
3-06-102111	Grant Coordinator of a research center	Education
1-04-100511	Postdoctoral Fellow	Sociology (Demography)
4-25-120511	Postdoctoral Fellow	Education (Learning Sciences)
3-14-113011	Postdoctoral Fellow	Architectural history
4-19-012012	Postdoctoral Fellow	History
2-13-111411	Professor	Sociology
5-11-103111	Professor	Sociology
1-02-100511	Professor	Anthropology (Archaeology)
4-18-121911	Researcher	Anthropology (Archaeology)

Table 2. Rank and academic discipline of the study participants

Participants expressed several different perspectives on the relationship of their data to the demands placed on them for scholarly production and teaching, as well as on their access to university services and appropriate training. None of the scholars interviewed during this study expressed satisfaction with their level of expertise in data management, and few had access to individuals who could provide knowledgeable guidance. On the contrary, most participants reported feeling adrift when establishing protocols for managing their data and added that they lacked the resources to determine best practices, let alone to implement them. Almost none of the scholars reported that data curation training was part of their graduate curriculum. Data management was typically discussed only in research methods courses and often only at a cursory level of detail in relation to methodological approaches and problems. The difficulties involved in the practical aspects of managing and preserving large amounts of research data were rarely addressed in these methods courses, and most researchers reported learning the necessary skills "on the job" via trial and error.

The transition to digital data collection has altered scholarly workflows. The greater ease of collecting data digitally has likely

increased the amount of data collected, particularly with certain types of documentation, such as digital images. Researchers need no longer stretch a limited supplies budget to cover the high cost of film and, without this restriction, may be less judicious with their documentation. In the case of digital media, it is better to collect data than to regret and leave curation decisions to some abstract future date. Additionally, analog data collection requires a significant investment of effort in data entry prior to the analysis phase. Depending on the size of the project and the funding available, data may be double or triple entered by a series of undergraduate and graduate research assistants, or they may be outsourced to data entry professionals for the well funded projects. Collecting data digitally eliminates this labor investment and shortens the lag between observation and analysis.

Although digital data collection offers certain efficiencies in moving from the observation to the analysis phase, the associated data management tasks are not easily delegated. Efficient entry of analog data does not require any specialized skills beyond keyboarding accuracy, while effective digital data management requires both expertise and labor continuity that is not readily found in a pool of transient research assistants. Thus, an additional burden of labor has shifted to the scholars themselves, and they are grappling with ways to balance the changes in research labor with increasing expectations for teaching performance. The following sections summarize the most salient themes that emerged from the participant interviews.

Research Context and Workflow

Perhaps one of the more complicated issues for data curation is the complex life cycle of research data and the idiosyncratic growth of research projects. Rarely does data collection take place within a discrete phase of a project (figure 1). In fact, researchers may develop data protocols before the project is funded and may then change the protocols in response to issues as they arise. Collaborators may also join the project and contribute data that were collected under different circumstances. It may not be until the active research phase that data collection is systematic, although changes in protocol may occur even during this phase. In some cases, a project does not work out as planned, and researchers recycle it into a new research idea or take it in a new direction entirely.

Additionally, scholars may collect data on a phenomenon unrelated to their current project with no clear idea of the potential usefulness of those data. Such data might be integrated with a later project, given away to an interested colleague, or never used at all. For example, Participant #2-12-111011, Assistant Professor, Environmental Studies collected data on graffiti during fieldwork and then donated the data to another researcher (see Appendix C, case study #3). It is perhaps unrealistic to expect that research will follow a well defined, linear progression that can be neatly categorized. Importantly, because the researchers themselves could not always predict

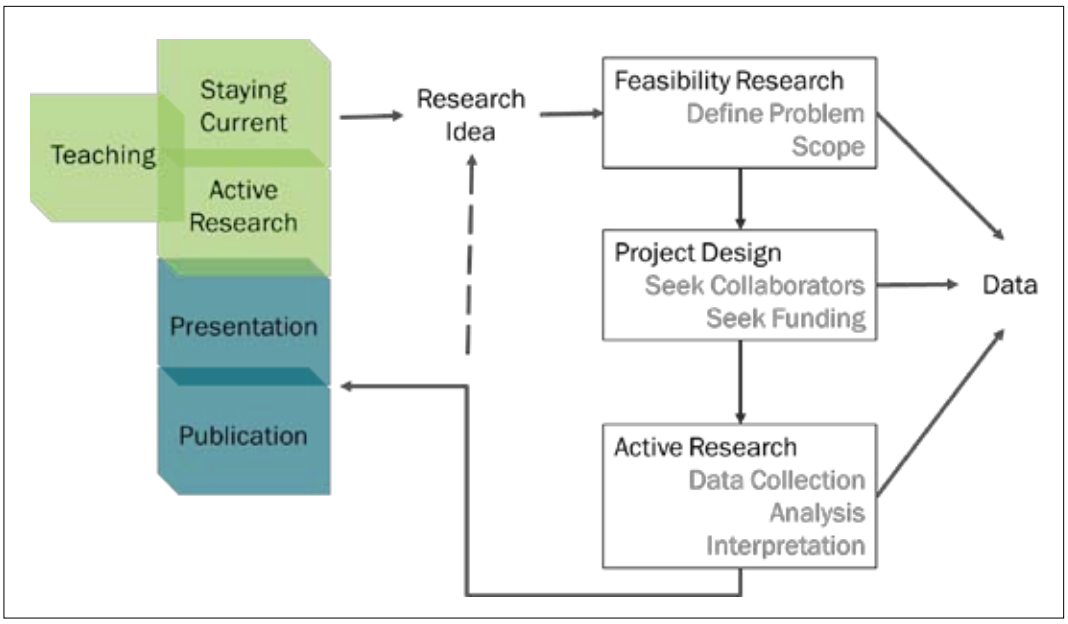


Fig. 1. Research workflow of a typical scholar showing the nonlinear development of research projects and the multiple stages at which data are collected

which data would be useful in the future (either for themselves or for other researchers), they were unsure which data should be preserved and what contextual information should be included with the data.

Several participants commented on the nonlinear nature of the research process and the way that this complicates data analysis and management. Participant #1-17-121211 described the nonlinearity of the research process as one of the most challenging aspects of working with data and noted the dynamic relationship between storage, analysis, and communicating results:

> It would be nice if there was a way to collect data, have it migrate into a collection space, and in the collection space get it prepared for whatever analytics you're going to engage in. And then have a place for the output of the analytics to go back into that collection space so that they're connected in some way [with] your analysis and the data you've collected, which makes it easier to engage in the process of writing or making sense of this. Where you're not simply looking at the results you also have access to the instruments that you used to collect and the questions that were related to those, as in the research questions, but also the instrument questions. Because those things have a way of finding their way into your. . . your write-up of the data or of your analysis but because of the sort of disparate and heterogeneous nature of it, becomes it's like, you know, chasing cats (Participant #1-17-121211, Assistant Professor, Education).

This participant went on to describe tools that could remediate some of these difficulties, suggesting networked databases that include tools for ingesting data according to schema designed for the project's research questions. Framing data ingestion with the

research questions would facilitate linking the research findings to the analysis and observations. Cokol and colleagues (2005) discuss similar ideas for integrating research questions with data to improve analysis. The logistics of implementing such a system aside, this participant's comments underscore the need for developing data management strategies early in the research process.

The researchers held contradictory views about the value of their data. Some study participants wondered who might be interested in their data while also expressing a desire to associate their data with publications or to have it available for use in the classroom (e.g., Participant #2-12-111011, Assistant Professor, Environmental Science). Other researchers wanted to create products other than research articles, such as websites, or to share certain aspects of the data, but they cited a lack of the skills or time needed to do so (Participant #3-06-102111, Grant Coordinator, Education).

Few of the researchers in this study thought about long-term preservation of their data, especially those who were early in their career. Perspectives regarding research data tend to be pragmatic. Given the nature of the academic system, which offers little or no career reward for preserving one's data, this is not surprising. Typically, metadata and documentation are of interest to researchers only if it helps them complete their work and produce publications. After a project ends, the time required to add appropriate metadata often exceeds the researcher's capacity and willingness to edit it, and the demands of publication output overwhelm long-term considerations of data curation. Many of the researchers were also skeptical of long-term interest in their data and were often doubtful that future researchers would be interested in their primary materials. This doubt contributes to scholars' reluctance to allocate time to data preservation and annotation. Scholars are in great need of basic archival skills to help them set priorities for data curation tasks and decide which data should be preserved.

Overall, the researchers interviewed for this study exhibited an extremely wide range of data collection practices and habits, and they readily adapted research workflows to fit their current interests and needs. For this reason, file formats, as well as the software and hardware platforms used to manage and manipulate data, tend to proliferate. Data preservation strategies not only must take into account these varied, proprietary, and non-standard data formats, but also must provide a real-time benefit for the scholar in meeting research goals.

Collaboration and Data Sharing

Researchers need better online collaboration tools that provide more sophisticated access controls and can support the volume of data generated. Participants frequently reported exceeding their data quotas within university networks, and they sought tools that allow them to collaborate across institutions and manage data in a networked environment. Consequently, they routinely resorted to

a constellation of personal computers, external hard drives, and commercial spaces, further compounding technical issues in data management. Several of the participants in this study were working on collaborative research projects that spanned multiple institutions (e.g., Participants #4-18-121911, #1-03-100511, #3-06-102111, and #1-17-121211), prompting project directors to seek non-university file-sharing options, such as Dropbox or Google Docs.

Using commercial "cloud" services as data storage locations poses potential privacy and security problems since the terms of service for these products are often poorly understood by researchers and the research participants. Furthermore, the terms of service may not be sufficient to meet the data protection and confidentiality standards that researchers and their institutional review boards (IRBs) require. Dropbox's well publicized June 2011 security glitch, which left all Dropbox accounts open to access without a password for several hours, is indicative of this problem. Applying additional security measures, such as encrypting files locally prior to sharing them via a cloud service, is beyond the technical skills of many researchers, and it diminishes the ease of use that leads researchers to adopt these tools as a file-sharing solution. Universities' common practice of limiting access to institutional networks to formally affiliated individuals has also contributed to this problem by making university-based systems of little use to multi-institutional collaborations. Universities should consider amending these policies to reflect the reality of multi-institutional research teams.

The field of physics offers a valuable lesson regarding the storage of data in personal accounts, as recounted by Curry (2011). From 1979 to 1986 a particle detector experiment called JADE (Japan, Deutschland, England) was performed at the PETRA e+e collider in Hamburg, Germany; the experiment resulted in several important discoveries for particle physics. In the more than 25 years since, theoretical insights and computing advancements have made the JADE data valuable once again. However, much of the data have been irrevocably lost to corrupt storage media, lost computer code, and deactivated personal accounts. These early particle physics experiments are unique, as modern colliders operate at higher energy levels and cannot replicate the particle interactions. Given the lack of infrastructure for sharing and storing data, the social sciences may face similar problems of data loss in documenting social phenomena as researchers begin to work within larger collaborative groups and with larger data sets. Data stored on personal media devices are especially vulnerable to this type of loss, as few scholars have the skills necessary to maintain data over time and across hardware and software platforms. Several of the scholars interviewed reported storing data on legacy systems that may become inaccessible (e.g., Participants #2-15-120211, #2-22-021512, #1-02-100511).

Although some researchers would welcome greater ease in sharing their data, particularly in collaborative projects, many are reluctant to enter into any arrangement in which they would relinquish control over access to the data. As one researcher described:

> [P]eople don't hesitate, at all, to share data with collaborators that they trust.… If you provide a mechanism for collaboration, even if it's just Google Docs or something, you know, people share data easily and freely. It's when it becomes an anonymous process that they seem to get balky (Participant #1-02-100511, Professor, Anthropology).

The willingness to share may be related to proximate goals in that easier data sharing facilitates collaboration within the project and reduces the proliferation of file versions, a routinely cited difficulty (e.g., Participants #2-16-120211, #1-17-121211, #3-06-102111).

Perspectives regarding data sharing beyond the research project are much more complex. Researchers have reported various ownership issues related to their data, and they are sensitive to the effects that releasing data might have on individuals related to the project (e.g., collections curators or study participants unintentionally identified). Researcher concerns related to protecting data privacy range from ensuring the physical safety of research participants (Participant #2-13-111411) to helping prevent the theft of objects from museums or other research locations (Participant #4-18-121911). Some researchers also report that ethical concerns about the appropriate use of their data underlie their desire to maintain control over who can access the data. Concerns regarding the misuse of data become particularly important in studies of marginalized groups of people and politically sensitive issues. Confidential and nonconfidential data are often intermingled in the data sets of social scientists, causing them to be inherently conservative about data sharing. In the following excerpt, a professor of sociology comments on the relationship of trends in social science data to the need for technological infrastructure that supports diligent privacy protection.

> Sheer size can be a problem, but clearly the biggest problem is the problem of protection of privacy. The concern of privacy has been ramped up tremendously over the last 10 or 15 years, and the process of getting permission to analyze data can be difficult, but a trend in social science data is to include more and more information that's sensitive. A lot of studies now include certain biomarkers and so a difficulty for us is providing secure facilities to do this, 'cause frequently now a national survey organization will require very restrictive conditions. So, I'm also the director of something called the [name omitted], which is an organization that spans both campuses, has about 50 faculty associates, and we have a number of people who are analyzing this kind of data. And we have actually set up our own "cold room" at the School of Public Health, and we were looking forward to the library actually setting up a "cold room" or a "cool room" for us when the new [name omitted] commons building opens to accommodate this. While I have some resources as a professor to do some of this on my own, graduate students don't that are working and in general, how to do this is a problem. I think the field is trying to now establish appropriate levels of protection

> for particular kinds of data and is trying to balance this problem of privacy with public access, and it is a challenge and it's going to require some new modes of doing things. In one of our projects at the population center we're partnering with ICPSR [Inter-university Consortium for Political and Social Research] to see if we can test distributed cold rooms where we would have a computer here at [university name omitted] that would have encoded communications, let's say, with the computer at ICPSR so that we would never have the data here. So managing this kind of restricted access is difficult, especially for social scientists when they don't have multimillion dollar grants. It's becoming a bigger and bigger issue as the data gets better and better, i.e., has DNA markers in it (Participant #5-11-103111, Professor, Sociology).

The protective attitude toward research data might also lead (or even require) researchers to neglect metadata and secondary materials (e.g., codebooks, explanatory materials, finding aids, ontologies) that are necessary to ensure the long-term usefulness of primary data. If data are not to be disseminated, these aids are often unnecessary to individuals or small groups of researchers. Platforms that could provide both a workspace and a preservation space would add significant value for scholars. Additionally, university policies that appropriately address the ethical considerations relating to data sharing and preservation would benefit researchers, administrators, and technologists alike. These policies must go beyond the determination of who has access to which equipment to address the changing relationship of information to electronic identity and its influence on individual rights.

Training, Technical Issues, and Infrastructure

None of the researchers interviewed for this study had received formal training in data management practices. They were learning on the job in an ad hoc fashion. A few of the participants had consulted with experts in the field (e.g., Participant #5-09-103111 had consulted the Smithsonian Institution for guidance regarding the preservation of 16-bit color raw files) or had used self-help books and syllabi found online (Participants #1-04-100511 and #5-09-103111). By far, the most common strategy was to apply lessons learned in theory and methodology courses (e.g., statistics) and then learn by trial and error. The best-case scenario encountered during this study was a project at Penn State University that emphasizes ontology development at the beginning of the research process. Thus, graduate students and junior researchers received some training in data practices specific to that project while working within the lab or project team.

Few of the researchers interviewed for this study had developed a long-term data management plan for their research data. In the case of those who had developed a plan, the requirements of an outside funding agency, such as the National Science Foundation, were often the motivating factor. Nevertheless, the variety of audiences

who might utilize the data (e.g., other scholars, policymakers, the public at large), as well as a lack of metadata standards for preserving information about a project, hindered researchers' efforts to effectively share and disseminate their data.

The researchers are not naïve; they understand that poor data management can be costly to their research and that access to greater technical expertise, through either a consultant or additional training, would be useful for their work. However, it is unlikely that many researchers would undertake additional training. Participants in this study repeatedly cited a lack of time to conduct basic organizational tasks, let alone time to research best practices or participate in training sessions.

> [S]o some organized system that is good for putting notes in but which you could easily attach files to would be good. Frankly, I've got so much stuff to do that I'm not likely to do that. Like I said, my guess is I could do that, my guess is you could attach, you could certainly attach links in a OneNote document, you might be able to attach the documents for all I know, but I also need someone to tell me that it's in my interest to do it and kind of prod me and help me do it. Both urge me and help me to do it at the same time. 'Cause otherwise, I'm not likely to archive stuff (Participant #5-11-103111, Professor, Sociology).

As with the creation of metadata, the economics of the scholarly reward system are likely to influence researcher perspectives on additional training (i.e., such training seems extraneous, as it does not directly contribute to publication production).

Researchers report that a variety of technical issues, such as inadequate access to networked storage, data loss because of poor organization, legacy file formats, and the scale of their data, can overwhelm available infrastructure. Although some of these issues stem from a lack of training or knowledge about best practices for data management, the issues cannot be separated from access to adequate infrastructure. As one researcher described:

> [O]ne of the things that's really helped us in the very recent past is being able to store all of our data or nearly all of our data on a server somewhere.... The infrastructure has to be there, I'm realizing now, in order to be able to even begin to organize yourself…it's a combination of, sort of people and, and hardware that has to be there in order to facilitate someone like me who has a lot of data being able to manage those data effectively (Participant #1-03-100511, Assistant Professor, Anthropology).

The participant went on to describe a colleague's more generously funded project that includes database programmers who manage large data sets of computed tomography (CT) scans. He emphasized the importance of having individuals who work closely with the project manage some of the technical aspects. This kind of support is beyond the means of most projects, leaving the researchers to manage data on their own. As another participant put it, "We really don't

have the level of expertise or the person dedicated to this that would bring, you know, the whole thing to fruition on the scale in which it's envisioned" (Participant #4-18-121911, Researcher, Anthropology). As a result of this gap in technical expertise, parts of the project were scaled back or suspended indefinitely.

Researchers hold tremendous amounts of data on personal computers and hard drives, many of which are not backed up adequately. Among the participants in this study, the scale of research data ranged from under 1 GB to multiple terabytes. Data types included various formats of images, video, audio files, data sets (public and original), documents (paper and digital), code packages, and analysis scripts. Even individuals who are early in their research career may have amassed significant bodies of data (e.g., Participant #3-14-113011, a postdoctoral fellow, already had thousands of image files). Managing large files presents significant challenges for researchers in that university infrastructures typically do not provide adequate storage space or sufficient bandwidth for data access (e.g., Participant #4-25-120511 could not store videos from interviews with study participants on university servers). These data are vulnerable to loss when researchers upgrade their computers or software, and few researchers put more than minimal effort into organizing non-active data or ensuring its continued compatibility with new software or hardware.

Role of the Library

There is a clear need for libraries to move beyond passively providing technology to embrace the changes in scholarly production that emerging technologies have brought. Few researchers see the library as a partner, and most of the researchers in this study seemed to regard the library as a dispensary of goods (i.e., books, articles) rather than a locus for badly needed, real-time professional support. However, Participant #5-11-103111 characterized the library as an ideal location to create spaces for working with restricted data in compliance with governmental and other guidelines. These spaces would be particularly useful for graduate students and junior faculty who may not have their own labs. Furthermore, the creation of such spaces could facilitate researcher integration with data preservation programs.

FINDINGS AND RECOMMENDATIONS

The data preservation step must be fully integrated into a scholar's research workflow. Not only are necessary metadata and other materials much more easily captured while research is in progress, but also there is a real opportunity to streamline research workflows and to provide much needed support. Scholars need help with the technical aspects of managing and preserving data, as well as with basic curation issues (e.g., what to keep and what to delete), and the ethical implications of sharing their data (e.g., what is an appropriate

latency period for the data and how does one balance the need to provide meaningful access with the risk of inadvertently exposing confidential participant information).

Although some researchers acknowledge that their data could be useful to other researchers, there is little incentive to invest time in archiving or repackaging data sets. In fact, investing time in a project beyond its usefulness for publication is counterproductive, given the high expectations for producing research publications. In such cases, reframing data curation within a comprehensive backup and management strategy is potentially valuable; for example, it may be helpful to point out that data curation contributes to the success of the ongoing research program by alleviating many of the technical issues researchers face (e.g., data loss caused by poor organization, version issues, management of obsolete file formats for long-term projects, and provision of secure collaboration tools). Arguments aimed at convincing researchers to think about long-term data preservation for its own sake are not likely to be effective.

Our findings and recommendations are as follows:

1. An approach that emphasizes early engagement with researchers and dialog around finding/building the appropriate tools to manage data for a particular project/researcher is likely to be the most productive.
   a. There is unlikely to be a single out-of-the-box solution that can be applied to the problem of data preservation.
   b. Extensive outreach to scholars is necessary to build the relationships that will facilitate data preservation. This is likely to be a slow process initially.
   c. Researchers are unlikely to engage with those they do not view as peers.
2. Researchers must have access to adequate networked storage.
   a. Universities should not restrict access to infrastructure to individuals in permanent faculty positions.
   b. Universities should revise their network policies to support multi-institutional research projects.
   c. Researchers need additional tools to manage preserved data on their own, and they would benefit from access to professionals who can offer advice on management strategies.
3. Improved privacy and data access control are needed.
   a. It is essential to develop tools that manage confidential data and provide the necessary security. Most importantly, policies must be developed that support researchers in the use of these technologies.
   b. These systems must ensure that researchers have control over their data, as well as over who has access to it. Without such assurances, many researchers are unlikely to invest in these systems. In many cases, their desire to avoid the ethical risks of inappropriate data release may outweigh the costs of potential data loss.
4. Early intervention in the researcher career path is likely to have the greatest benefit.

a. Working with graduate students as they develop their first major research project is a key opportunity for education in best practices and the importance of good data management protocols.
b. Young scholars often have not considered the long-term value of their data or the importance of a systematic approach to data management. As their research develops and they begin teaching, they are likely to regret neglecting data management.
c. Small- to medium-sized research teams and single researchers are likely to have the greatest unmet need, because they typically lack the resources of major research initiatives to hire data professionals.

Researchers typically align themselves with their disciplines rather than with their institutions; therefore, support models that extend beyond the university are likely to be especially beneficial. Scholars also spend substantial periods of their careers migrating among institutions, particularly during the early phases. Researchers who are in temporary positions may not be willing to commit to a university data management system when they may leave in a year or two, and they may fear that they will be unable to retrieve their data at that time. Furthermore, research projects are frequently both interdisciplinary and interinstitutional. Thus, systems that restrict access to institutional affiliates would preclude multi-institutional collaboration among scholars in data sharing and preservation.

Reaching the level of collaboration among universities and the technical interoperability required to capture and preserve a career's worth of data in the current environment is a challenge. A practical model for fostering both collaboration and interoperability may be a network of local data specialists who are aligned with disciplines and/or affiliated with a regional or national scholarly organization. A local data specialist who operates within the university to collaborate with researchers and who participates in a network that extends beyond the university would facilitate long-term collaboration with researchers as they move through the various stages of their career. Such a network would also provide the communication necessary to foster interoperability in technical solutions.

Our interviews with researchers suggest that data specialists should have at least some expertise—preferably considerable knowledge—in the discipline with which they are working. In the best-case scenario, a data specialist would be fully integrated into a research team and would also conduct research. These specialists are likely to need significant technical training in addition to their subject knowledge. However, given the variation of research modalities and the types of data generated, it is difficult to ascertain what type of technical training they will need until they are on the job. An iterative approach to training that builds on core technical skills and emphasizes identification of needs specific to subject or methodological areas may be effective.

Finally, it is likely that a data specialist will need to function as an advocate for researchers within the local systems. Although some

universities already provide technological or other support that would be useful for researchers, the bureaucracy surrounding this support can severely limit researcher access. Scholars may not have the time or knowledge necessary to influence the policies that affect them. Furthermore, researchers of junior rank may not have sufficient influence to affect relevant policies. Thus, some basic training in policy development, negotiation, and academic administration may be useful for data specialists.

CONCLUSION

Current data management systems must be fundamentally improved so that they can meet the capacity demand for secure storage and transmission of research data. Integrating the data preservation system with the active research cycle is essential to encourage researcher investment. Enhancing the system with intuitive live linking visualization tools could add significant value for preliminary analyses (Fox and Hendler 2011), as well as curatorial decision-making.

There is also a clear need for "privacy enhanced protocols" (both policy and technical) that address the ethical concerns of researchers while creating standards for data latency, access, and attribution (Altman and King 2007; King 2011; Lawrence, Jones, and Matthews 2011). Researchers are not well positioned to meet the technical and policy challenges without the coordinated support of libraries, information technology units, and professionals who possess both technical and research expertise.

REFERENCES

Akil, Huda, Maryann E. Martone, and David C. Van Essen. 2011. Challenges and Opportunities in Mining Neuroscience Data. *Science* 331(6018): 708–712.

Altman, Micah, and Gary King. 2007. A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-lib Magazine* 13(3/4): 1–13.

Arnett, Jeffrey J. 2008. The Neglected 95%: Why American Psychology Needs to Become Less American. *The American Psychologist* 63(7): 602-614.

Carnegie Foundation for the Advancement of Teaching. 2010. *Carnegie Classification of Institutions of Higher Education*. Available at http://classifications.carnegiefoundation.org/.

Cokol, Murat, Ivan Iossifov, Chani Weinreb, and Andrey Rzhetsky. 2005. Emergent Behavior of Growing Knowledge About Molecular Interactions. *Nature Biotechnology* 23(10): 1243–1247.

Curry, Andrew. 2011. Rescue of Old Data Offers Lesson for Particle Physicists. *Science* 331(6018): 694.

Evans, James A., and Jacob G. Foster. 2011. Metaknowledge. *Science* 331(6018): 721–725.

Fox, Peter, and James Hendler. 2011. Changing the Equation on Scientific Data Visualization. *Science* 331(6018): 705–708.

Gur, Ruben C., Farzin Irani, Sarah Seligman, et al. 2011. Challenges and Opportunities for Genomic Developmental Neuropsychology: Examples from the Penn-Drexel Collaborative Battery. *The Clinical Neuropsychologist* 25(6): 1029–1041.

Harris, Mark. 2007. *Ways of Knowing: Anthropological Approaches to Crafting Experience and Knowledge*. Brooklyn, NY: Berghahn Books.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. The Weirdest People in the World? *The Behavioral and Brain Sciences* 33(2-3): 61–83; discussion 83–135.

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

King, Gary. 2011. Ensuring the Data-rich Future of the Social Sciences. *Science* 331(6018): 719–721.

Lang, Trudie. 2011. Advancing Global Health Research Through Digital Technology and Sharing Data. *Science* 331(6018): 714–717.

Lawrence, Bryan, Catherine Jones, and Brian Matthews. 2011. Citation and Peer Review of Data: Moving Towards Formal Data Publication. *The International Journal of Digital Curation* 6(2): 4–37.

Mathews, Debra J. H., Gregory D. Graff, Krishanu Saha, and David E. Winickoff. 2011. Access to Stem Cells and Data: Persons, Property Rights, and Scientific Progress. *Science* 331(6018): 725–727.

Nisbett, Richard E. 2003. *The Geography of Thought: How Asians and Westerners Think Differently—and Why*. New York: Free Press.

Overpeck, Jonathan T., Gerald A. Meehl, Sandrine Bony, and David R. Easterling. 2011. Climate Data Challenges in the 21st Century. *Science* 331(6018): 700–702.

Pool, Ithiel de Sola. 1983. Tracking the Flow of Information. *Science* 221(4611): 609–613.

Rzhetsky, Andrey, Ivan Iossifov, Ji Meng Loh, and Kevin P. White. 2006. Microparadigms: Chains of Collective Reasoning in Publications About Molecular Interactions. *Proceedings of the National Academy of Sciences of the United States of America* 103(13): 4940–4945.

Smail, Daniel Lord. 2008. *On Deep History and the Brain*. Berkeley and Los Angeles: University of California Press.

Appendix A: Data Overview

Research Activities/Example Projects

- Imaging of primate bone morphology and walking behavior in juvenile humans
- Girls' expectations and transition to adulthood
- Cognitive development in children, using eye tracking equipment
- Behavioral diagnostics for educational programs and support for special needs program students
- Criminal justice policy analysis
- Reanalysis of archeological excavation site data
- Secondary analysis of literature
- Learning among children within an online environment
- Environmental issues and political protest in Kyrgyzstan
- Community-based nongovernmental organizations and civil society in Ethiopia
- Architectural history and landscape (Europe)
- Decision-making among Indian prime ministers: The policymaking process
- Effect of notebook computers on foreign language teachers
- Indian legal history and the British Empire
- Transformation of the welfare system in Turkey and the relationship to grassroots politics
- Changes in U.S. welfare policy, 1990–2006 (multisite study of 2,500 low-income families)
- Archaeological tourism in Highland Bolivia, 2002–2004
- Mummy bundles, Peru (historical archeological research), Database integrating data (between institutions)
- Development of a prototype for digital curation microservices (tools/applications driven by services, e.g., ingest of object/authenticating object, version control)
- Data curation for Antarctica McMurdo Dry Valleys (18 years of data): Documenting the magmatic plumbing system
- Antiterror laws in Turkey, prosecution of the Kurdish minority
- Archeology in the Gordian region, Turkey (and collaboration with civil rights nongovernmental organization)

Reported Issues

Collaboration

- Management of workflow across multiple campuses
- Inadequate online collaboration space
- Inadequate tools to manage versioning, etc.

Infrastructure

- Systems and infrastructure overwhelmed by scale of data
- Policy (e.g., varying levels of access complicate workflows for research teams that include undergraduate and graduate students)

Data loss

- Parts of personal archives lost (e.g., computer crash, organizational mistakes)

- Inadequate time and skill to maintain data in legacy file formats (e.g., MS Word)

Data sharing

- Lack of systems to adequately segregate and maintain control over sensitive data
- Some data considered proprietary by collection holders (museum collections)
- Philosophical perspectives on data sharing (e.g., ethical considerations, methodological complexities)
- Lack of suitable mechanisms for sharing

Training, support, and personal organization

- Little or no training, learning as needed throughout research
- No contact with university data services
- No archival planning
- Very limited backup procedures
- Difficulty maintaining and tracking support materials
- Unclear about need for (and definition of) metadata
- Unsure of best practices regarding preservation in terms of file formats
- Difficulty deciding what should be preserved and what should be destroyed
- Difficulty maintaining organizational structure of files, insufficient time for organizational tasks

Common Data Types

- **Images**: TIFF, raw, JPEG, KML (for display of geographic data)
- **Video**: mp4, mov
- **Audio**: wav, mp3, analog tape
- **Data files**: Excel, SPSS, STATA, ArcGIS, txt, various public data sets
- **Documents**: MS Word, PDF
- **Paper-based**: Manuscripts, newspapers, site reports, transcripts, field notes (often handwritten notebooks, sometimes scanned, but rarely transcribed), drawings/sketches, chemical analysis results, photographs
- **Other**: Code packages and documentation, tool prototypes, objects (artifacts/samples), Matlab scripts (e.g., Participant # 3-05-102111 uses Matlab scripts to transform txt files for analysis in SPSS)

Analytical Tools

- Google Earth
- ArcGIS
- ProfilesPlus (specialized software for policy analysis)
- Excel
- SPSS
- STATA
- Matlab

- Atlas.ti
- Qualtrics
- NVivo,
- Filemaker Pro
- MS Access

Management Tools

- File structure on personal computer and naming conventions
- Excel
- E-mail
- Website
- OneNote

Data Storage

The volume of data varies from a few gigabytes or smaller to multiple terabytes. Researchers report storing data in a variety of locations, including:

- University server system (RAID)
- Inter-university Consortium for Political and Social Research (ICPSR) archive (Participant #5-11-103111)
- SharePoint
- Personal computers (usually multiple)
- Work computers
- External hard drive
- "Cloud" storage (e.g., Google Docs, Dropbox)
- DVD
- Office (for physical materials)

Study participants are using a variety of locations to store data and are employing many combinations of the various locations. In some cases, they are using multiple locations because the capacity of any one location is insufficient to support the volume of data while enabling access from multiple locations (e.g., terabyte scale data of Participant #1-03-100511). In other cases, the dispersal of data reflects idiosyncratic work habits with insufficient time for organizational tasks.

Collaboration Tools

- File sharing software: Dropbox, Google Docs, SharePoint
- Database programs: Bento by Filemaker Pro, MS Access, Filemaker
- Communication and record keeping: Wikis (university and non-university), e-mail (university and non-university), Skype, other conference calling
- Hardware: University network, networked drives (within the lab), flash drives
- Outsourcing to a data support company

Collaboration Problems

- Versioning issues
- Volume of data too large for university networks (e.g., Participant #1-03-100511 had to mail a hard drive)
- Uneven access to university infrastructure (e.g., Participant #4-25-120511 reported that undergraduates and graduate students on a project do not have the same privileges as senior project members for network storage)

Appendix B: Interview Questions

Demographics:

- What is your academic discipline?
- What is your position?
- When did you complete your highest academic degree?
- Did your graduate program include training in curating or managing data?
- How would you define digital curation?

Background:

- Ask the participant to describe a research project she/he is currently working on (or recently completed). Ask the participant to narrate the process of completing the work from beginning to end.
- What were the goals of this project?
- How did you become involved in this project?
- What kind of data sources did you use in this project?
- What kinds of primary sources did you use?
- What kinds of secondary sources did you use?
- How did you locate these data sources?
- Did this project have a data preservation or a data management plan requirement?

Data Creation/Analysis:

- Did you create new data sources as part of this research (e.g., experimental results, data sets, coding files, indexes)? What kind?
- When/how were these data collected? Is data collection still active? If so, when do you expect it to be completed?
- What are the formats of the data used in this project?
- How many items are contained in the data set?
- How large are the files?
- How did you organize the data?
- How are the data named/numbered, etc.?
- Did you document this system?
- Are the data backed up? How/Where?
- How do you work with/analyze/manipulate/transform the data?
- What tools do you use? What formats do you work with?
- What problems have you encountered while working with the data?
- What are the products/outcomes of your work?

Collaboration:

- Do you collaborate with other researchers on this project?
- How do you manage this collaboration?
- How did you manage version control?
- What software (if any) did you use?
- If you wanted to go back and work with the data again, what would be the most important information to have?
- If someone wanted to replicate/reconstruct your analysis, what information would be needed?

Preservation:

- Once you were finished with this project, what happened to your research materials/data?
- Where are they located? In what format?
- Did anyone offer guidance in making these decisions?
- Do you have a plan/strategy for archiving these materials?
- Where will they be held?
- Who is responsible for them?
- (If not) Why don't you archive your materials?
- What concerns do you have about archiving or curating your data?
- If someone were to return to your data in 5 to 10 years (or longer), what contextual information would be needed?
- If you were archiving your research for future scholars, what would be the most important things to be preserved?
- Who would potentially re-use this data?
- What are your expectations for this re-use (e.g., citation, copies of papers, reciprocity)?
- Do your data contain confidential and/or proprietary information (e.g., personally identifiable information, patentable information)?
- Would you publish your original data if you believed there was a suitable venue?
- What concerns do you have regarding publication methods?
- What are the most important factors when deciding if data are suitable for publication?
- Does your university or library offer any services to help you with curating your data?
- If the university (or library) were to offer services to help you with data curation, what would be the most helpful things they could provide?

Personal Practices and Training:

- Do you keep a personal archive of materials related to your scholarship (e.g., field notes, lab books, e-mails, photographs)?
- What formats are these materials in?
- How/Where are they stored?
- Have you had training in data curation?
- If so, what kind/what tools?
- Who provided the training?
- Do you feel that it was adequate?
- What would you like to know more about?
- During what phase of your research development did you receive this training?
- Did the timing seem appropriate for your work?
- How did the training influence the way you conducted your research?

For individuals fulfilling the role of digital curator:

- Do you conduct outreach as part of your curator responsibilities? If not, does another staff member fulfill this role?
- Who is the primary audience for outreach?
- Have the efforts been successful in engaging faculty or other stakeholders?
- What would you change about this process?

Appendix C: Case Studies

Case Study #1: Data Curation for the Antarctica McMurdo Dry Valleys Project

Participant #5-09-103111 was the only scholar interviewed during this study who was working in a position that was formally designated as a digital curator (in this case, a "data scientist"). Nevertheless, this researcher had no formal training in data curation except for his attendance at a summer institute at the University of Illinois. However, the researcher holds a master's degree in both computer science and geology, giving him the combination of technical skills and deep disciplinary knowledge that is necessary for managing the data of the complex project he described.

This scholar's project is the digital preservation and curation of approximately 18 years of research materials and geologic data collected in the McMurdo Dry Valleys of Antarctica. The data are diverse, including both physical and digital artifacts, and his tenure has spanned the migration of data collection from analog to "born digital" formats. At the outset of the project, the data curator made a detailed catalog of all data in need of preservation and noted the difficulty of archiving the materials in an electronic form. The materials to be archived include researchers' field notes, personal journals of field seasons, chemical analyses, maps and aerial photographs, photographs (about 4,500 35-mm slides, as well as other images in a range of digital formats), geologic samples, thin sections (cut sections of rock mounted on glass slides to be viewed via microscope), and video of fluid dynamics experiments.

The goal of this project was to preserve and present as much of the material online as possible, and several types of materials presented particular difficulties. Analog 35-mm slides had to be converted to digital formats using a specialized Nikon slide scanner (Coolscan 5000; 16-bit color). The data curator consulted with the Smithsonian Institution for format preservation guidance and decided on an uncompressed TIFF format at the highest resolution available for long-tem preservation and JPEG files at lower resolution for presentation purposes. Associating sufficient metadata (including location and date) with photographs was often problematic, as the research team had included little or no metadata with the original photographs; some important photographs require time-consuming annotation by the original researcher. Excel spreadsheets were used to track the necessary metadata for the image files.

Physical objects have also proved difficult to present online. To obtain high-quality images of the geologic rock samples (more than 800), it was necessary to contract with a professional photographer. The thin sections also posed difficulties, because the images needed enough resolution to allow researchers to measure 200–500 grains of the mineral. Pixilation on lower resolution images renders them unusable, making very large files (up to 60–65 GB per section) necessary. These files not only create storage problems, since up to

100–150 TB are needed for the project, but also require specialized software tools to make the images usable online.

This project is currently in progress, and the team envisions a wide range of potential audiences for the curated materials, including other researchers, the general public, and primary and secondary students. The digital curator said that while the data have been prepared thus far principally for other researchers and therefore require an understanding of geological fieldwork to be meaningful, he envisions an "interactive geologic map" that would be useful to a wide audience.

Case Study #2: Walking Behavior in Juvenile Humans

Participant #1-03-100511 is a biological anthropologist who studies primate evolution and primate bone morphology using image data (high-resolution computed tomography). He is presently an assistant professor (doctorate completed in 2001) and had no digital curation or data management training as part of his graduate training. This scholar's current project is a National Science Foundation (NSF)–funded, multi-institutional study of bone development and its relationship to the walking behavior of juvenile humans.

Data for this project are initially collected in an imaging lab and then processed locally in the researcher's anthropology lab. Digital image files are transferred from an acquisition computer to a server, where they are maintained and backed up. The workflow for processing the bone images for analysis is complex and requires multiple specialized software programs for three-dimensional visualization and measurement. There are several thousand TIFF images for a single bone, and images are repositioned, sampled, and extracted to a Digital Imaging and Communications in Medicine (DICOM) format so that measurements can be made. The numerical data are analyzed in SPSS and Excel.

Tracking metadata for the images as they pass through the multiple processing steps has proved difficult. The initial bone imaging data include XML files with the metadata describing the scanner settings. The researchers wrote a custom PERL script to extract the metadata required for analysis as a text file, which is then imported into an Excel spreadsheet for tracking purposes. The researcher organizes and manages project data using a Windows file structure. However, metadata are not always held at every level of the file structure, and the members of the research team must consult the tracking spreadsheet, which sometimes creates confusion.

The need to share files among researchers at multiple universities has also created problems. FTP sites and university server solutions failed for technical reasons, requiring the researchers to mail a hard drive to members of the research team at another university. Tracking and metadata files have been shared via Dropbox, which initially created conflicting copies of documents and required the design of new workflows to avoid duplication.

Although this project has both an NSF data management plan

and a physical anthropology data-sharing plan (a standard in physical anthropology for a number of years), several factors limit the effective reuse of the project's research data. The ultimate goal for this project is to maintain all data sets indefinitely and potentially to make these data available for download via a website. However, no database is currently in place to make this possible, and the volume of materials (terabyte scale) makes preparing a database and the necessary metadata time-consuming; scanner settings must be described to reproduce the researcher's methodology.

Bone collections often have tight restrictions on their use and reuse. For example, when the project needed chimpanzee bones to use for comparison with human bones, the researchers could not obtain samples locally. Thus, the researcher had to travel to Belgium to use a collection there, resulting in scans made on different types of equipment that required different processing steps. In addition, collection owners (e.g., museums) may consider bone scans proprietary, and they may assert ownership over data produced from their collections, limiting the sharing of data. In this case, data rights can become a source of conflict, as the researcher's institution asserts ownership over data produced by university-owned scanners. These conflicts over data ownership and rights effectively render data unusable for other researchers and can lead data managers to be very conservative in their sharing processing and practices.

Case Study #3: Environmental Issues and Political Protest in Kyrgyzstan

Participant #2-12-111011 is an assistant professor of environmental science who studies environmental politics and protests in Kyrgyzstan.

This scholar collects quantitative and qualitative data using face-to-face interviews, as well as secondary data sets. She holds interview data on paper questionnaire forms, as well as in audio recordings. Quantitative results are stored in Excel and SPSS files, while the audio recordings are in the process of being transcribed. The researcher hopes to scan the print versions of her questionnaire forms and destroy the originals, which are presently stored in boxes in her office. Because she works in three languages (Kyrgyz, Russian, and English), the researcher has had difficulties hiring and training transcriptionists, and the transcription of her interviews has taken several years to complete. Transcription files have been managed by means of flash drives and Google Docs.

The researcher is concerned about her skills in data management. Although she has significant experience working with secondary data sets, she has had no formal training in data curation. In particular, she observed that she has a weak and nonsystematic backup plan for her data, relying principally on multiple personal computers and external hard drives. The organization of digital files is also very difficult for this researcher, and she finds the file management tools

that are part of a computer's operating system insufficient for her needs.

None of this researcher's funding agencies have required a data-sharing or data management plan. However, she has a vague plan for preserving and making public her data, and she hopes to make some of her data available for use by other scholars and policymakers, particularly the quantitative data sets that she has used to conduct spatial analysis in geographic information systems (GIS). She is also interested in the potential for making public her qualitative interview results and notes, but has concerns about confidentiality and privacy. The researcher hopes to maintain her materials indefinitely for her own use—preferably on a university server (she is presently doing this for her GIS data).

This researcher's experience demonstrates the unexpected and unpredictable uses of data sets. A graduate student working on graffiti images following an ethnic conflict in Kyrgyzstan asked the researcher for permission to use copies of photographs that the researcher had taken and cataloged in the immediate aftermath of a particular event. The researcher had taken the photos purely out of interest, and they were not directly relevant to her current research or future plans. Notably, the researcher also holds an electronic collection of Kyrgyz newspapers that no longer exist and no longer have web archives.

Data Curation Education: A Snapshot

Spencer D. C. Keralis

"We believe professionals in all fields need a richer understanding of how their professions and the materials they work with are being transformed by the emergence of the digital information ecosystem."

—Peter Boticelli et al., "Educating Digital Curators: Challenges and Opportunities"

This study provides a snapshot of the current digital data curation education landscape. Because the field is rapidly changing in response to several factors—an increasingly demanding job market, the needs of researchers who must cope with data management planning mandates from national funding agencies, and the perceived "data deluge" that threatens to overwhelm the research and library communities in terms of technology, infrastructure, and staffing—this snapshot is necessarily limited in scope and marks a specific moment in time.

The study has three main goals:

1. To describe how library and information science (LIS) programs address digital data curation as a component of their curricula for librarians
2. To describe the extra-academic training curricula developed by scholars and professionals to address unmet needs within their communities
3. To use this information to make recommendations for training curriculum development for future CLIR fellows

For the purpose of this discussion, digital data curation is best described as life cycle data management; it encompasses a spectrum of activities ranging from research data management planning at the project inception stage; through collection of data as part of the

research process; through the identification, processing, and accession of data sets; and, finally, to the archival preservation and sharing of data in an appropriate repository. The term *data* in this context refers to "*everything* needed to have reproducible science" (Woods Hole Oceanographic Institution 2012). Although in the present discussion these concepts are concerned primarily with the sciences and social sciences, they are applicable across disciplines for any research that relies on or generates data.

DATA CURATION IN THE LIS FIELD

Those in the LIS field perceive data curation as an intrinsic part of their discipline. Data curation education efforts are most often embedded in standard LIS courses (for example, as components or modules of metadata and database architecture courses), and efforts to teach data curation as a discrete set of intelligible practices are both recent and few. Currently, only five LIS schools offer graduate certificates explicitly in data curation. These tracks are part of programs that lead to a master's degree in library and information science (MLIS), with the certificate requirements distributed over the progression of the two-year program, and are generally not open to non-LIS students or professionals.

These programs, isolated within the standard LIS curriculum or within certificate programs that are exclusive to LIS students, are not designed to meet the needs of researchers or professionals who may benefit from these skills. Furthermore, researchers' perception of libraries as "a dispensary of goods … rather than a locus for real-time research/professional support" compromises the ability of those in the LIS field to intervene effectively in campus research activities and may even foreclose collaboration with other disciplines (Jahnke and Asher 2012, 4).[1] As Weber and associates note in their report on the 2010 Data Curation Research Summit, "LIS will need to develop stronger partnerships with domain researchers, informaticists, and other stakeholders in the research enterprise, to succeed at making research data an integral and enduring part of the information assets retained for science and scholarship over the long term" (Weber et al. 2011, 6).

The most valuable intervention to come out of the LIS field for the purposes of digital data curation education is the development of a matrix of skills and functions by Cal Lee at the University of North Carolina at Chapel Hill. The DigCCurr Matrix describes 24 functional areas and 4 meta-level functions (Lee 2009). These are broad, high-level categories, designed to address "digital curation 'know how,' as opposed to the conceptual, attitudinal or declarative knowledge." Defining these skills potentially makes it possible to develop a modular, skills-based curriculum that can be customized for different skill levels and functional concentrations.

[1] Although this perception is a commonplace complaint among academic librarians, the anthropological portion of this project may well be the first time this has been formally documented as a phenomenon, and may merit further study.

Research conducted by Virgil Varvel and associates at the University of Illinois at Urbana-Champaign as part of the Data Conservancy project demonstrates the difficulty of identifying data curation tracks within LIS curricula. Using a keyword search based on concepts in the DigCCurr Matrix to survey "online course catalogs and websites of 63 iSchools and other LIS schools,"these researchers uncovered "475 courses in 158 programs at 55 schools" (Varvel, Bammerlin, and Palmer 2012). The net cast by this project was wide, as the researchers included introductory LIS courses containing foundational knowledge that may be developed in later courses (although the results published thus far do not indicate whether the researchers attempted to make such connections between courses to see if this was borne out within individual curricula) and "exceptions were made if information was ambiguous, to err on the side of inclusion" (528).

The study broke out four categories of courses:

1. Data-centric—"courses were focused exclusively on data curation, data management, or data science topics" (8 percent)
2. Data-inclusive—"courses have segments devoted to data topics related to e-science or e-research" (11 percent)
3. Digital—"courses *did not appear to explicitly attend to research data expertise*, they included digital topics that are highly relevant for education of data professionals" [emphasis added] such as "digital library development" or "digital preservation or digital collections and services" (27 percent)
4. Traditional LIS—courses that "give students an introduction to important topics developed further in data inclusive or data centric courses" (54 percent)

The Data Curation Curriculum Search tool developed through the research of Varvel and associates does not allow a search based on these categories, and these categories do not appear as descriptors in individual course records within the tool. As a result, it is impossible with the information available publicly to provide examples of each for further examination.

Data Conservancy researchers claim that the percentage distribution among the course categories "indicat[es] a high level of coverage of at least *some aspects* of data expertise" [emphasis added]. However, more than half of the courses identified in the study are "traditional LIS"—the most ambiguous category and the one that the researchers allowed themselves to most "err on the side of inclusion." More than one-quarter of the courses identified fall into the digital category, but while these courses include skills that may in some ways be transferable to the data curation environment, they do not explicitly address the needs of data-intensive research. Thus, 81 percent of the courses identified require some evaluation before they can become part of a curriculum for data curation professionals, while less than 10 percent are specific to the state of education in data curation.

Given the apparent improbability that students will encounter a data-centric course in their line of study, it seems that students must

already be well versed enough in the language of data and the needs of researchers to evaluate course descriptions, must be committed to constructing a data-intensive education for themselves, or must have an advisor knowledgeable enough to help them craft a track from traditional LIS courses in order to come out of most existing U.S. LIS programs with the skills and knowledge necessary to support the needs of data-intensive research.

Current Data Curation Certificate Programs

The United Kingdom's Digital Curation Centre (2012a) identifies five data management certification programs in the United States (table 1). Each of these programs restricts its enrollment to LIS students, with the exception of the University of Arizona's DigIn! Program, which admits post-baccalaureate students and professionals who are not enrolled in Arizona's MLIS program.

Institution	Program	Mode	URL
University of Arizona	Graduate Certificate in Digital Information Management*	Distance	http://digin.arizona.edu/
University of California at Berkeley	Master of Information Management and Systems	Residential	http://www.ischool.berkeley.edu/programs/masters
University of Illinois at Urbana-Champaign	Data Curation Education Program (DCEP)	Residential	http://www.lis.uiuc.edu/programs/ms/data_curation.html
University of North Carolina at Chapel Hill	DigCCurrI (master's students); DigCCurrII (doctoral students)	Residential	http://ils.unc.edu/digccurr/institute.html
San Jose State University	Master's Degree in Archives and Records Administration (MARA)	Distance	http://slisweb.sjsu.edu/mara/

Table 1. Data management certification programs in the United States

* The development of the University of Arizona program is described by Peter Botticelli et al. (2011).

Varvel and associates, in their research for the Data Conservancy, identify a larger pool of certifications that may be applicable to data curation (Varvel, Bammerlin, and Palmer 2012). They "identified 7 master's degree programs, 4 certificate programs, and 10 other concentrations with a specific emphasis on data in their descriptions at 17 different institutions." However, they point out that some of these programs are data-in-name-only: Even though they have "data" in their descriptions, they included few data-centric or data-inclusive courses—a fact that seems to undercut the optimism expressed by the researchers about the potential for these programs to produce data professionals. (Unfortunately, Varvel and associates do not call out these programs by name.)

Emerging Data Curation Certificate Programs

There are several digital curation certificate programs under development at institutions around the United States, but two programs are of particular interest to this study.

The first is at the University of North Texas iSchool, which is developing a Graduate Academic Certificate in Digital Curation and Data Management. This program will be open to non-LIS students and to non-student professionals from the sciences and social sciences, computer science, and the humanities, as well as to LIS master's and doctoral students. The curriculum will be a modular grouping of non-residential online courses, but will require onsite capstone sessions with LIS faculty. A pilot version of the initial course, Cyberinfrastructure Fundamentals for Digital Curation and Data Management, will launch in the summer of 2012 (University of North Texas 2011).

The second program of interest is a partnership between the Purdue University Libraries and the libraries of Cornell University, the University of Minnesota, and the University of Oregon. This program will "develop a training program in data information literacy for graduate students who will become the next generation of scientists." At each institution, teams of librarians and experienced researchers will develop "a shareable data information literacy training curriculum for students in science/engineering graduate programs" (Institute of Museum and Library Services 2011). The outcomes of these parallel development efforts will be evaluated and shared online for the use of other libraries.

In the emerging programs identified so far, the trends are toward allowing open enrollment for scholars and professionals outside the LIS discipline and toward developing more collaborative models of teaching and learning that partner librarians and LIS educators with research faculty. In some cases, the digital data curation certificate program is not based in the LIS school at all; at the University of Maine, for example, the New Media Studies program will host the interdisciplinary Digital Curation Graduate Certificate. Museum studies programs are also beginning to offer digital curation certificates that address the specific needs of museums in identifying, preserving, and providing access to digital artifacts, born-digital art, and other assets (Pratt Institute 2012).[2] Table 2 includes a few of the certificate programs under development; this record is far from comprehensive, however. As of the 2011 funding cycle, the Institute of Museum and Library Services (IMLS) had awarded more than $9 million to data curation education and capacity building, indicating a commitment to developing data expertise further in LIS professionals.

[2] For more on digital curation curricula in museum studies, see Tibbo and Duff (2008).

Institution	Program	Funder	Launch Date	Enrollment
Pratt Institute	Project CHART! (Cultural Heritage Access Research and Technology)	IMLS	Fall 2012	LIS only
Purdue University	Next Generation Scientists	IMLS	Fall 2012	Open
University of Maine	Digital Curation Graduate Certificate	Unknown	Fall 2012	Open
University of North Texas	Graduate Academic Certificate in Digital Curation and Data Management	IMLS	Pilot begins Summer 2012	Open

Table 2. Sample data curation certificate programs under development

Other emerging educational efforts in data curation do not involve an academic certificate. Rather, they move toward embedding LIS students in research environments. In 2010, the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign Illinois received a $988,543 Laura Bush 21st Century Librarian Program grant from IMLS to develop "a sustainable and transferable model for educating library and information science master's and doctoral students in data curation through field experience in research and data centers." The Data Curation Education in Research Centers (DCERC) program involves a partnership between the GSLIS, the National Center for Atmospheric Research (NCAR), and the University of Tennessee, School of Information Sciences. This model is valuable in that it embeds students in research and data centers, but the program is open only to enrolled master's and doctoral students in the iSchool at Illinois and the University of Tennessee, School of Information Sciences.

EXTRA-ACADEMIC TRAINING PROGRAMS

Several extra-academic programs provide potential models for training postdoctoral scholars in digital data curation. Some of these programs originated in the efforts of LIS schools to address the needs of professionals, while others have emerged from groups of professionals seeking to fill in the gaps in their training and to build communities of practitioners with similar interests and needs.

DigCCurr II Professional Institutes

The DigCCurr program at the University of North Carolina at Chapel Hill offers annual professional institutes "aimed at assisting digital collection managers in developing their digital curation strategies" (DigCCurr 2012). The program began in 2009 and has been held every year since then. Each institute includes a spring program with a winter follow-up session and public symposium.

Digital Preservation Outreach and Education

The mission of Digital Preservation Outreach and Education (DPOE), an initiative of the Library of Congress, is "to foster national outreach and education to encourage individuals and organizations to actively preserve their digital content, building on a collaborative network of instructors, contributors, and institutional partners" (DPOE 2012).

From September 20-23, 2011, the DPOE Baseline Train-the-Trainer workshop was held at the Library of Congress. Developed in partnership with Nancy Y. McGovern of the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, the DPOE Train-the-Trainer Workshop for digital data preservation provides attendees with a basic digital data preservation curriculum, as well as with "tips and techniques for conducting successful workshops." The workshop consists of six modules:

1. Identify: What digital content do you have?
2. Select: What portion of that content is it your responsibility to preserve?
3. Store: How should digital content be stored for the long term?
4. Protect: What steps need to be taken to protect your digital content?
5. Provide: How should digital content be made available?
6. Manage: What provisions should be made for long-term management?

Digital data preservation educators teach the workshops. Graduates of the program are able to offer the workshops at their home institutions for researchers and practitioners within their region. The focus is on preservation rather than life cycle data management, and participants are expected to have a fairly significant technical background prior to participating in the workshop.

Digital Curation Centre

Describing the organization as "the UK's leading hub of expertise in curating digital research data," the Digital Curation Centre (DCC) website is a clearinghouse of information for practitioners seeking advice or resources on data management. The DCC also offers workshops in data management, including Data Curation 101, a three-day intensive course for data custodians. For beginners, DC101 Lite distills the information in DC101 into a half-day course. The courses are structured around the DCC Curation Lifecycle Model 1. Unlike the DPOE model, which focuses on preservation, the DCC model addresses the full range of issues in digital data curation (DCC 2012b). The course materials are available online to share and reuse.

The DCC also offers a train-the-trainer program, which makes the generic DC 101 and DC 101 Lite training materials available for use "as the basis for disciplinary or institutional-specific training."

CURATEcamp

CURATEcamp is a series of "unconference" events for digital data curation practitioners. The camps deliberately include a wide range of practitioners, recognizing that "digital curation is a practice that happens all over: libraries, archives, public media, industry, start-ups, non-profits, government, and so forth."

The attendees at these unconferences set the agenda of each camp, though often with a pre-described theme or concept in mind. For example, the October 2011 CURATEcamp that occurred in conjunction with the Digital Library Federation (DLF) Forum had the theme "Catalogers and Coders" and brought together metadata specialists and technologists "to engage in interactive problem solving and exploration of topics of joint interest, especially in the area of Linked Data."

Although CURATECamps are no doubt useful as forums for the exchange of information and ideas, perhaps their most valuable function is the creation of diverse communities of practitioners who are confronting similar issues in a wide range of disciplines.

A Note on Certification

Each of these extra-academic training models offers its participants an opportunity to develop particular skills and knowledge, and in some cases, participation carries a certain cachet for those familiar with the programs. Institutional alignment can also convey credibility; for example, the DPOE program bears the imprimatur of the Library of Congress. However, none of the models can deliver industry standard or academically recognized accreditation or certification. Participants can supplement their experience in these programs with software or other industry certifications, but accreditation and certification would be the strongest incentives for participants to invest the time and make the financial commitment required for academic programs.

CONCLUSIONS

Although the IMLS is investing heavily in data-oriented education in the LIS field, and LIS and iSchool programs are making efforts to develop data curation curricula, much work still needs to be done to prepare LIS graduates for roles as data professionals in and out of libraries. Furthermore, the LIS world largely remains a closed circuit, providing concentrations within tracks restricted to LIS enrollees. The trend in emerging curriculum development programs is to open up this closed circuit and allow post-baccalaureate students and professionals to take courses in data curation; this trend can only strengthen the LIS programs and those professionals taking part in them. Data curation is not a single-discipline practice, and developing programs that include professionals and students from across the natural, social, computer, and information sciences, and the

humanities will help produce practitioners who are better prepared to meet the needs of data-intensive research.

The Council on Library and Information Resources (CLIR) Postdoctoral Fellowship in Academic Libraries is a proven model for preparing doctoral scholars for service in academic libraries. CLIR's weeklong "library bootcamp" introduces fellows to some of the issues facing twenty-first century libraries, creates a cohort of fellows who can share experiences and information, and helps realign the newly minted Ph.D.s in relation to the academy. Host institutions benefit from library-friendly scholars who are able to work intensively on both service and research initiatives within the libraries.

In 2012, the DLF program of CLIR received a $679,827 grant from the Alfred P. Sloan Foundation to help launch the new CLIR/DLF Data Curation Fellowship Program. The program, an expansion of CLIR's Postdoctoral Fellowships in Academic Libraries, will provide recent Ph.D.s with professional development, education, and training opportunities in data curation for the natural and social sciences. For these fellows, the CLIR bootcamp model will be expanded and adapted to include an additional skills-based practicum that will introduce fellows to the terminology, tools, and issues they will face in their positions. Library and LIS professionals will be recruited to provide the training.

The experience gained during the two-year postdoctoral fellowships will encourage the development of highly skilled and knowledgeable specialists. The aim is to create a cadre of scholarly practitioners who understand not only the nature and processes of their own disciplines but also the ways in which their research data are organized, transmitted, and manipulated. For the program's first cohort, CLIR is now recruiting six data curation fellows in cooperation with its partner institutions: Indiana University, Lehigh University, McMaster University, Purdue University, the University of California at Los Angeles, and the University of Michigan.

Bibliography and Links

Botticelli, Peter, et al. 2011. Educating Digital Curators: Challenges and Opportunities. *The International Journal of Digital Curation* 2(6).

Council on Library and Information Resources. "CLIR/DLF Data Curation Postdoctoral Fellowship." Available at http://www.clir.org/fellowships/datacuration.

CURATEcamp. Available at http://curatecamp.org/.

DigCCurr. 2012. "About DigCCurr II: Extending an International Digital Curation Curriculum to Doctoral Students and Practitioners." Available at http://ils.unc.edu/digccurr/aboutII.html#institutes.

Digital Curation Centre. 2012a. "Data management courses and training." Available at http://www.dcc.ac.uk/node/8975.

Digital Curation Centre. 2012b. "Curation Lifecycle Model." Available at http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf.

Digital Preservation Outreach and Education (DPOE). Available at http://www.digitalpreservation.gov/education/.

Institute of Museum and Library Services. 2011. "National Leadership Grants–September 2011 Grant Announcement." Available at http://www.imls.gov/news/national_leadership_grant_announcement.aspx#IN.

Jahnke, Lori, and Andrew Asher. 2012. The Problem of Data: Data Management and Curation Practices Among University Researchers. *The Problem of Data*, pp. 3–31. Washington, DC: Council on Library and Information Resources.

Lee, Christopher. 2009. "Functions and Skills (Dimension 2 of Matrix of Digital Curation Knowledge and Competencies)." June 18, 2009 (Version 18) Project: DigCCurr (IMLS Grant # RE-05-06-0044). School of Information and Library Science, University of North Carolina at Chapel Hill. Available at http://www.ils.unc.edu/digccurr/digccurr-functions.html.

Pratt Institute. 2012. "Project CHART." Available at http://www.pratt.edu/academics/information_and_library_sciences/grant_scholarship_internship/chart/.

San Jose State University. "Master's Degree in Archives and Records Administration (MARA)." Available at http://slisweb.sjsu.edu/mara/.

Tibbo, Helen, and Wendy Duff. 2008. "Toward a Digital Curation

Curriculum for Museum Studies: A North American Perspective." 2008 Annual Conference of CIDOC. Athens, September 15–18, 2008. Available at http://cidoc.mediahost.org/archive/cidoc2008/Documents/papers/drfile.2008-06-23.pdf.

University of Arizona. "Graduate Certificate in Digital Information Management." Available at http://digin.arizona.edu/.

University of California at Berkeley. "Master of Information Management and Systems." Available at http://www.ischool.berkeley.edu/programs/masters.

University of Illinois at Urbana-Champaign. "Data Curation Education Program (DCEP)." Available at http://www.lis.uiuc.edu/programs/ms/data_curation.html.

University of Maine. "Coming Soon: Digital Curation Graduate Certificate." Available at http://umaineonline.umaine.edu/certificate/certificates-coming-soon-spring-2012/digital-curation/.

University of Maine. "Digital Curation at the University of Maine." Available at http://digitalcuration.umaine.edu/.

University of North Texas. 2011. "Library, College of Information only recipients of multiple research awards." UNT In-House (August 8). Available at http://inhouse.unt.edu/library-college-information-only-recipients-multiple-research-awards.

Varvel, Virgil E. Jr., Elin J. Bammerlin, and Carole L. Palmer. 2012. Education for Data Professionals: A Study of Current Courses and Programs. *Proceedings of the 2012 iConference*, 527–529. New York: Association for Computing Machinery.

Weber, Nicholas, Tiffany Chao, Carole L. Palmer, and Virgil E. Varvel Jr. 2011. *Report on the Data Curation Research Summit.* Paper presented during the 6th International Digital Curation Conference, December 6, 2010, Chicago. Champaign, IL: Center for Informatics Research in Science & Scholarship, University of Illinois. Available at https://www.ideals.illinois.edu/handle/2142/28355.

Woods Hole Oceanographic Institution. 2012. "Data Management and Publishing." Available at http://www.whoi.edu/DoR/page.do?pid=44235.

Other Resources

Bailey, Charles Jr. *Digital Curation and Preservation Bibliography 2010*. Available at http://digital-scholarship.org/dcpb/dcpb2010.htm

Data Curation Curriculum Search. Available at http://cirssweb.lis.illinois.edu/DCCourseScan1/index.html.

DigCurV: Digital Curator Vocational Education Europe–A registry of courses and workshops for digital curators in Europe. Available at http://www.digcur-education.org/eng.

The Digital Curation Exchange (DCE**).** Launched at the DigCCurr 2009 conference, DCE is a central location for digital curation professionals, educators, and students to discuss "all things digital curation." Available at www.digitalcurationexchange.org.

DPOE Training Calendar**.** The first national calendar of digital preservation training opportunities in the United States. Available at http://www.digitalpreservation.gov/education/courses/index.html.