

VOLUME ONE

# CURATING RESEARCH DATA

*Practical Strategies for Your Digital Repository*



EDITED BY LISA R. JOHNSTON



# Curating Research Data

Volume One: Practical  
Strategies for Your Digital  
Repository

*edited by*  
*Lisa R. Johnston*

*Association of College and Research Libraries*  
*A division of the American Library Association*  
*Chicago, Illinois 2017*

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1992. ∞

Cataloging-in-Publication data is on file with the Library of Congress

Copyright ©2017 by the Association of College and Research Libraries.  
All rights reserved except those which may be granted by Sections 107 and 108 of the Copyright Revision Act of 1976.

Printed in the United States of America.

21 20 19 18 17 5 4 3 2 1

Cover image

Copyright: kentoh / 123RF Stock Photo ([http://www.123rf.com/profile\\_kentoh](http://www.123rf.com/profile_kentoh))

# Table of Contents

## 1 .....Introduction to Data Curation

*Lisa R. Johnston*

Data, Data Repositories, and Data Curation: Our Terminology

Why We Curate Research Data

The Challenge of Providing Data Curation Services

Reuse: the Ultimate Goal of Data Curation?

Conclusion

Notes

Bibliography

## Part I. Setting the Stage for Data Curation. Policies, Culture, and Collaboration

## 33 .....Chapter 1. Research and the Changing Nature of Data Repositories

*Karen S. Baker and Ruth E. Duerr*

Introduction

Background

*Changing Support for Data*

*Expanding Support for Data in Natural and Social Sciences*

*Data Repository Diversity*

Three Concepts at Work

*Data Ecosystem: Growing Interdependence*

*Liaison Work and Mediation*

*Continuing Design: Standards, Systems, and Models*

Changing Research Needs and New Initiatives

Final Thoughts

Notes

Bibliography

## 61 .....Chapter 2. Institutional, Funder, and Journal Data Policies

*Kristin Briney, Abigail Goben, and Lisa Zilinski*

Funding Agency Data Policies

Institutional Data Policies

Journal Data Policies

Navigating the Data Policy Landscape for Curation

Summary

Notes

Bibliography

**79 .....Chapter 3. Collaborative Research Data Curation Services: A View from Canada**

*Eugene Barsky, Larry Laliberté, Amber Leahey, and Leanne Trimble*

**Canadian Academic Library Involvement in Research Data**

**Management**

**Overview of Case Studies**

*Local Services: University of Alberta Libraries*

*Informal Regional Consortia: University of British Columbia Library*

*Formal Regional Consortia: The Ontario Council of University Libraries*

**Data Repository Services in Canadian Libraries**

*Discovery and Access Platforms*

*Long-Term Preservation*

**Operational Costs of Data Repository Services**

**National Collaboration: Portage**

*Goal 1: Portage National Data Preservation Infrastructure*

*Goal 2: Portage Network of Expertise*

**Future Directions**

**Conclusions**

**Notes**

**Bibliography**

**103 .....Chapter 4. Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation**

*Ixchel M. Faniel and Elizabeth Yakel*

**Introduction**

**Overview and Methodology for the DIPIR Project**

**Disciplinary Traditions for Data Sharing and Reuse**

*Social Scientists*

*Archaeologists*

*Zoologists*

**Data Reuse and Trust**

*Trust Marker: Data Producer*

*Trust Marker: Documentation*

*Trust Marker: Publications and Prior Reuse Indicators*

*Trust Marker: Repository Reputation*

**Sources of Additional Support for Data Reuse**

*Social Scientists*

*Archaeologists*

*Zoologists*

**Implications for Repository Practice**

**Conclusion**

**Acknowledgments**

**Notes**

**Bibliography**

**127 .....Chapter 5. Overlooked and Overrated Data Sharing:  
Why Some Scientists Are Confused and/or Dismissive**

*Heidi J. Imker*

Data Sharing in Context

*Overlooked Data Sharing: Article Publication*

*Overlooked Data Sharing: Supplemental Material*

*Overrated Data Sharing: Unsustained Community Resources*

*Overrated Data Sharing: Hyperbolic Arguments*

Conclusions

Acknowledgments

Notes

Bibliography

**Part II. Data Curation Services in Action**

**153 .....Chapter 6. Research Data Services Maturity in  
Academic Libraries**

*Inna Kouper, Kathleen Fear, Mayu Ishida, Christine Kollen, and  
Sarah C. Williams*

Introduction

Research Data and Libraries

The Current Landscape

RDS Maturity

Looking into the Future

Appendix 6A: Typology of Services and Their Descriptions on Websites

Notes

Bibliography

**171 .....Chapter 7. Extending Data Curation Service Models for  
Academic Library and Institutional Repositories**

*Jon Wheeler*

Introduction

Conceptual Models and Rationale

Alignment with Existing Roles and Capabilities

Applications: Requirements and Example Use Cases

Defining Stakeholder Interactions and Requirements

Harvesting and Metadata Processing

Content Curation and Packaging

Conclusion

Acknowledgments

Notes

Bibliography

**193 .....Chapter 8. Beyond Cost Recovery: Revenue Models and  
Practices for Data Repositories in Academia**

*Karl Nilsen*

Introduction

**From Costs to Revenue**

**Data Repository Revenue Models**

*Model 1: Public or Consortium*

*Model 2: Freemium*

*Model 3: Pay-to-Play*

*Model 4: Pay-if-You-Can or Pay-if-You-Want*

*Model 5: Grants*

*Model 6: Outside-Data*

**Common Challenges Associated with Revenue Practices**

**Conclusion**

**Notes**

**Bibliography**

**213 .....Chapter 9. Current Outreach and Marketing Practices  
for Research Data Repositories**

*Katherine J. Gerwig*

**The Survey**

**The Interviews**

*Measuring the Success of Repository Promotions*

*Successful Promotional Techniques*

*Unsuccessful Promotional Techniques*

*Target Audiences*

*Challenges to Increasing Awareness*

*Differences in Promoting the Institutional Repository and the Data  
Repository*

*Looking for Inspiration*

**Discussion**

**Conclusion**

**Promotional Examples for Inspiration**

**Acknowledgments**

**Appendix 9A: Data Repository Promotional Practices—Initial Google  
Survey**

**Notes**

**Bibliography**

**Part III. Preparing Data for the Future. Ethical and  
Appropriate Reuse of Data**

**235 .....Chapter 10. Open Exit: Reaching the End of the Data  
Life Cycle**

*Andrea Ogier, Natsuko Nicholls, and Ryan Speer*

**Introduction**

**Comparative Exploration**

*"End of Life Cycle" Terminology*

*Scope*

*Authority*

*Appraisal Criteria*

*Resources (Human, Financial, and Spatial)*

**Discussion***University Records and Information Management**Library Collections**Data Curation***Conclusion****Notes****Bibliography****251 ..... Chapter 11. The Current State of Meta-Repositories for Data***Cynthia R. Hudson Vitale***Introduction****Community Initiatives and Solutions to Support Meta-Repositories of Data****Methods****256 Results***Content**Functionality**Metadata***Discussion****Conclusion****Notes****Bibliography****263 ..... Chapter 12. Curation of Scientific Data at Risk of Loss: Data Rescue and Dissemination***Robert R. Downs and Robert S. Chen***Benefits of Data Rescue****Challenges of Data Rescue for Repositories****Repository Considerations for Data Rescue****Rescue of the Millennium Ecosystem Assessment (MA) Data****Dissemination of the Millennium Ecosystem Assessment (MA) Data****Lessons Learned****Discussion and Conclusion****Acknowledgments****Notes****Bibliography****279 ..... Contributor Biographies****Editor Biography****Author Biographies**







INTRODUCTION TO VOLUME ONE

# Introduction to Data Curation

*Lisa R. Johnston*

As varied as they can be rare and precious, data are becoming the proverbial coin of the digital realm: a research commodity that might purchase reputation credit in a disciplinary culture of data sharing or buy transparency when faced with funding agency mandates or publisher scrutiny. Unlike most monetary systems, however, digital data can flow in all too great abundance. Not only does this currency actually “grow” on trees, but it comes from animals, books, thoughts, and each of us! And that is what makes data curation so essential. The abundance of digital research data challenges library and information science professionals to harness this flow of information streaming from research discovery and scholarly pursuit and preserve the unique evidence for future use. Our expertise as curators can help ensure the resiliency of digital data, and the information it represents, by addressing how the meaning, integrity, and provenance of digital data generated by researchers today will be captured and conveyed to future researchers over time.

The focus of *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository* and the companion *Volume Two: A Handbook of Current Practice* is to present those tasked with long-term stewardship of digital research data a blueprint for how to curate data for eventual reuse. There are many motivations for storing and preserving data, but the ultimate goal of reuse by others will be a theme for all that follows. Following a brief overview to the terminology used in the two volumes, this introduction will explore the external motivations that impact why we develop data curation services and the driving forces behind why researchers share their data, including federal data management requirements, publisher policies for data sharing, and an overall sea change of disciplinary expectations for digital data exchange. Next, this chapter will dive into some of the

challenges that practitioners in the library and archival fields face when curating digital research data as well as some emerging solutions. In closing we will explore the sea change stemming from data reuse, from the disruptive effects that data transparency and the reproducibility movement have had on the scholarly communication life cycle to the potentially democratizing effect of digital data availability worldwide.

## Data, Data Repositories, and Data Curation: Our Terminology

Data is an evolving term. At its core, data can be any information that is factual and can be analyzed. Data is “information in numerical form that can be digitally transmitted or processed.” But in the research setting, data can be more abstract and consist of any information object (numerical or otherwise).<sup>1</sup> For information science professionals, the term ‘research data’ has been recently defined as:

“data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results.... Research data may be experimental, observational, operational, data from a third party, from the public sector, monitoring data, processed data, or repurposed data.

Data are defined in the Digital Curation Center (DCC) Curation Lifecycle Model as “any information in the binary digital form” and is treated there in the sense of any digital information that be taken in a broad perspective.<sup>3</sup> Harvey describes the breadth of data as encompassing all things digital, based on the UNESCO’s Guidelines for the Preservation of Digital Heritage and takes into account the more subtle nuances of NSF’s description of “scientific data” to create a list of data objects to include:

- Data sets: Observational, computational, simulated, or otherwise recorded output
- Digital collections: A grouping of digital objects, such as a photo archive or a vast text-based library of digitized books, can be interpreted as one data set
- Learning objects: Videos, digital online tutorials
- Multimedia: Recordings of film, music, and performance art
- Software: Applications including the code and documentation files<sup>4</sup>

Sometimes primarily associated with the sciences, data can be found in any discipline and in many forms.<sup>5</sup> Data may be raw (e.g., numbers collected by an instrument), aggregated from multiple sources, or the product of a model, simulation, or visualization (e.g., a graphic or video). Digital humanities data might include digitized or born-digital texts and monographs, digital image libraries, and 3D models, such as those used for historic reconstruction of ancient or mythological sites.<sup>6</sup> Social scientists produce large quantities of data, including survey data and observational data, such as complex human activity and interactions captured via sensors or video.<sup>7</sup> Outside of research, the business, industry, and commerce sectors produce “big data” that is used to better understand research questions about human behavior, and as a result a growing (and sometimes nefarious) economy of selling the transactional data derived from business has emerged.<sup>8</sup>

With the explosion of digital data produced by modern research or recorded through our general day-to-day activity, digital data repositories are storing vast amounts of information. Data repositories preserve information “by taking ownership of the records, ensuring that they are understandable to the accessing community, and managing them so as to preserve their information content and Authenticity.”<sup>9</sup> The co-authors of the “Key Components of Data Publishing” report use the practitioner-based Research Data Alliance (RDA) definitions developed by the Data Foundations and Terminology Working Group and the Research Data Canada’s Glossary of Terms and Definitions to define digital repositories as:

A repository (also referred to as a data repository or digital data repository) is a searchable and queryable interfacing entity that is able to store, manage, maintain and curate Data/Digital Objects. A repository is a managed location (destination, directory or ‘bucket’) where digital data objects are registered, permanently stored, made accessible and retrievable, and curated. Repositories preserve, manage, and provide access to many types of digital material in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control for the digital material to be authentic, reliable, accessible and usable on a continuing basis.<sup>10</sup>

Additionally, the 2005 National Science Board anticipated the need for data repositories, stating that:

It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ush-

ered in such a fundamental change. Digital data collections are at the heart of this change. They enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. Through their very size and complexity, such digital collections provide new phenomena for study. At the same time, such collections are a powerful force for inclusion, removing barriers to participation at all ages and levels of education.<sup>11</sup>

Simply put: data includes a wide range of information, and data repositories retain this information for reuse. Therefore our challenge as data curators is to apply the archival principles of library and information sciences to a wide-variety of complex data objects from all disciplines and prepare them for ingest, access, and long-term preservation within an environment (such as a data repository) that facilitates discovery and access while not diminishing their context, authenticity, and value. No short order. As data curators we effectively become the first users of the data. In doing so we may review the various aspects of the data (such as arrangement, completeness, clarity, and quality), identify any reuse issues early on, and work with the data author to correct these issues. This concept is very important considering the long-term burden of ingesting and storing research data in our repositories. We need to first verify that those data can be understood and do our best to *optimize* them for reuse. Otherwise, our data repository can still do all of the things listed in the RDA definition above, the only difference being that the data might not be usable.

It is the variety and complexity of data, and its context, that make it much more difficult to preserve so that others might make use of it. Therefore our definition of data curation must also include verifying that all of the essential metadata and supplementary information, describing what the data is and how to understand it, are curated as well. For example, ensuring that supplementary files to the dataset, like codebooks, data dictionaries, schemas, and readme files provide the additional documentation needed to understand the file contents is a key step in the data curation process.

The optimization aspect can be found in the “adds values” statement of the University of Illinois’ School of Information Sciences Data Curation Specialization definition for data curation as

the active and ongoing management of data through its life-cycle of interest and usefulness to scholarship, science, and education. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation.<sup>12</sup>

However these concepts also apply to any digital object (for example, a book or an article), not necessarily just data, and therefore data curation is understood as a subset of digital curation which covers all types of digital information.<sup>13</sup> In short, the goal of data curation is to prepare research outputs in ways that make it useful beyond its original purpose, ensure completeness, and facilitate long-term citability.

Volume One of *Curating Research Data* explores the variety of reasons, motivations, and drivers for why data curation services are needed in the context of academic and disciplinary data repository efforts. The following twelve chapters, divided into three parts, take an in-depth look at the complex practice of data curation as it emerges around us. Part I sets the stage for data curation by describing current policies, data sharing cultures, and collaborative efforts underway that impact potential services. Part II brings several key issues, such as cost recovery and marketing strategy, into focus for practitioners when considering how to put data curation services into action. Finally, Part III describes the full life cycle of data by examining the ethical and practical reuse issues that data curation practitioners must consider as we strive to prepare data for the future.

## Why We Curate Research Data

In Part I, *Setting the Stage for Data Curation: Policies, Culture and Collaboration*, we explore the factors that influence our actions to provide data curation services for research data. Some factors include incentives, both scholarly positive and negative, from the funding bodies and the scholarly publishing entities. Other factors come directly from the research communities themselves, some of which are demanding greater transparency in research. These motivations can sometimes be indirect or at even at odds with a researcher's goals.<sup>14</sup> Overall the policies, culture, and collaborations involved with data curation provide us with an interesting canvas with which to begin our work.

One driving force that leads library and information science practitioners to provide data curation services is the inherent fact that digital data are more easily shared. Data have always held value beyond their original purpose, and today, digital data can travel and reach worldwide audiences at unprecedented speeds with incremental costs. A 1989 National Academies of Sciences panel described the impact of information technology on research in the sciences, engineering, and clinical research as improving collaboration among researchers "more widely and efficiently" by reducing "the constraints of speed, cost, and distance from the researcher."<sup>15</sup> And incentives to collaborate across institutional or disciplinary boundaries have boomed. Rates of co-authorship are increasing not only in the sciences but across disciplines that were traditionally solo-researcher focused such as the social sciences.<sup>16</sup> In short, digital data presents researchers with many new

ways of working collaboratively across institutional and geographic boundaries. **In Chapter 1, “Research and the Changing Nature of Data Repositories,” Karen S. Baker and Ruth E. Duerr draw from their experiences working at large scientific data repositories to explore data management and curation in the broader landscape of disciplinary research.** They describe how repositories, which initially were designed for highly structured data housed at key disciplinary repositories, have now emerged at the center of a modern ‘data ecosystem’ proliferated by the emerging requirements to openly, and ethically, disseminate research data. Their examples of early data registries and international data organizations—and the various stakeholders involved—paint a complex picture and provide excellent food for thought as our authors ask us to ponder how library data professionals contribute to and coordinate with the broader ecosystem of data repositories.

Another significant, and more opaque, driver for data curation services are the emerging funding requirements for data sharing. Over the last several years, national funding agencies and political administrations worldwide have developed a growing awareness of and the need for public access to the results of government-funded research and the long-term preservation of these unique digital research data sets.<sup>17</sup> For example, a key turning point in the US was the February 22, 2013 memorandum<sup>18</sup> by the White House Office of Science and Technology Policy (OSTP) directing federal agencies to develop plans to ensure all resulting publications and research data are publically accessible. The memo’s requirements for sharing digital research data in ways that make the data “publicly accessible to search, retrieve, and analyze” suggested that federally funded researchers will soon be faced with many new requirements that:

- Ensure that the data are richly described with machine-actionable metadata
- Ensure that data are complete, self-explanatory, and accurate (quality)
- Protect confidentiality and privacy when making data available (e.g., remove identifiers, virtual data enclaves)
- Account for the long-term access and preservation needs that go beyond the life of a grant.
- Identify and/or create trusted digital repositories to steward data over time<sup>19</sup>

Three years after the OSTP directive, “policies to make data and publications resulting from federally funded research publicly accessible are becoming the norm.”<sup>20</sup> Interestingly these efforts for sharing nationally funded research data run parallel to an open data movement for government-authored data. This movement is characterized by the G8 adoption of the “Open Data Charter” in June 2013 and demonstrated by the principles set forth in the US Open Data Action Plan released in 2014.<sup>21</sup> And not only federal funders that have moved the

needle towards open. Private funders of research, such as the Ford Foundation, the Alfred P. Sloan Foundation, and the Bill & Melinda Gates Foundation, now require their funded projects release underlying data with some degree of openness.<sup>22</sup> For a detailed listing of the current policies of federal agency responses to the OSTP memo, see SPARC Open Data's resource for Research Funder Data Sharing Policies.<sup>23</sup>

Complex? Absolutely. **Fortunately, Chapter 2, titled “Institutional, Funder, and Journal Data Policies” by Kristin Briney, Abigail Gobin, and Lisa D. Zilinski, does an excellent job of describing the current landscape of funder mandates for data as well as other top-down drivers for curation services.** For example, in 2009 the National Academies of Sciences put out a call for better standards for data sharing in ways that support reproducibility through the ethical sharing of data along with published research results. Authors of this report included editors of scientific journals that cited the emerging problem of “misguided efforts to clarify results” by distorting, falsifying, or even faking data.<sup>24</sup> This trend continues today and sources such as Retraction Watch regularly report examples of publishers responding to data-related issues in publications.<sup>25</sup> As a result, many journals have implemented policies to make the underlying data for an article more open to replication and validation. According to several studies such as Fear, Piwowar & Chapman, and Naughton & Kernohan of the Jisc-funded Journal of Research Data policy bank (JoRD) project, journal data sharing requirements come in many forms.<sup>26</sup> The latter in particular, after reviewing the data policies of nearly 400 journals, found that half did not have a data sharing policy and of those that did, 76 percent were found to be weakly worded and vague. In response the JoRD project developed a model data sharing policy that could be implemented by any organization.<sup>27</sup> Some prominent examples of journal data sharing policies include *Nature*, where “authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.” The *PLOS* data sharing policy goes one step further to say “Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection.”<sup>28</sup> Indeed, one such retraction occurred in 2015, albeit in a different journal (*Frontiers in Neuroscience*), due to an author refusing to share their data.<sup>29</sup>

Going beyond publisher requirements to simply make data accessible and linked to the article (see for example Elsevier's platform for linking data in data repositories such as PANGAEA), some publishers have created new journals that provide a venue for “data papers” or the long-form description of a dataset in conjunction with the data release.<sup>30</sup> Examples include Springer-Nature's *Scientific Data* and Elsevier's *Data in Brief* that both launched in 2014. The latter reports “an exponential rise in data articles over the six quarters since the journal came into existence, with approximately 300 publications expected in 2016 Q1.”<sup>31</sup> An independent survey of 116 data journals found that



the growth in data papers nearly doubled from 2012 to 2013 and continues to rise at an incredible rate.<sup>32</sup> Yet, one of the curious aspects of data journals is that the data are often not provided by the journal but rather “[the publisher does] not consider the publication of data as part of their own mission.”<sup>33</sup> For example, *Scientific Data* suggests a list of recommended data repositories for deposit since “we do not ourselves host data. Instead, we ask authors to submit datasets to an appropriate public data repository.”<sup>34</sup> It seems that scholarly communication is still rapidly adjusting to the new norm of data sharing and our data curation services will directly provide authors with the much-needed support.

International collaborations providing incentives for data curation services might be key. In 2004, many countries from Europe and others such as Australia, the US, and Canada signed the “Declaration on Access to Research Data from Public Funding” by the Organisation for Economic Co-operation and Development’s (OECDs) Committee for Scientific and Technological Policy, which set the stage for open access to digital research data resulting from public funding.<sup>35</sup> The results stemming from this Declaration have been substantial. In the United Kingdom, the seven councils of the Research Council UK (RCUK) and the private funder, the Wellcome Trust, have each established a policy on access to data in the years following the RCUKs 2011 report on “Common Principles on Data Policy.”<sup>36</sup> The European Commission has established a pilot program for data sharing through its Horizon 2020 granting arm.<sup>37</sup> And Canada’s three federal granting agencies are moving toward policies for research data such as those explored by Shearer in the comprehensive 2011 “Brief on Open Access to Publications and Research Data.”<sup>38</sup> **In Chapter 3, “Collaborative Research Data Curation Services: A View from Canada,” Eugene Barsky, Larry Laliberté, Amber Leahey, and Leanne Trimble provide in-depth case studies from their respective institutions, the University of British Columbia, the University of Alberta, and the Scholars Portal for the Ontario Council of University Libraries.** The three case studies are presented in the context of Canada’s overarching national infrastructure initiative, the ambitious Portage network developed by the Canadian Association of Research Libraries (CARL).<sup>39</sup> An exciting collaborative project, Portage aims to integrate existing research data repositories within a robust national discovery and preservation infrastructure network for all Canadian research data. Moreover the project will bring together library-based experts in order to share data management consultation services across a broader network. This national effort appears similar to the role that the JISC has played in the UK with its Research Data Management Shared Service Project and, on a much smaller scale for sharing curation staff expertise across institutions, the Data Curation Network project that your editor recently helped launch in the US in 2016.<sup>40</sup>

In Chapter 4, different disciplinary and cultural norms of how data reuse are explored by Ixchel M. Faniel and Elizabeth Yakel, who draw from ethnographic research with archaeologists, quantitative social scientists, and zoologists in “Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation.” To synthesize disciplinary data sharing and reuse findings the authors partner with three repositories—the Inter-university Consortium for Social and Political Research (ICPSR), Open Context, and the University of Michigan Museum of Zoology (UMMZ)—to obtain data reuse stories and even download statistics. Their study reveals the dependencies between how data are shared and how data are reused with emphasis on the differences in disciplines, and explores the interesting elements of “trust” in the data exchanged.

In Chapter 5, “Overlooked and Overrated Data Sharing: Why So Many Scientists are Confused and/or Dismissive,” Heidi J. Imker aptly focuses our attention away from scientists not or wrongly sharing their data to how often scientists share their data, and have historically been sharing data long before public access requirements. This chapter presents the idea that traditional methods of data sharing, though not generally meant for preservation purposes, are still valid forms of sharing within the discipline. For example, sharing data via publication in the traditional journal article is still very common, though much of this data is often fixed in graphs or charts found in the body of the article and therefore impractical or labor-intensive to reuse.<sup>41</sup> As one blogger quips, “Send me your data—pdf is fine,” said no one ever.”<sup>42</sup> Similarly, lengthy data tables historically induced costly page fees and data supplements to journal articles have been criticized as unstable and “far harder to locate than [data] in public repositories.”<sup>43</sup> Other widespread data sharing approaches, such as posting data to a project website or sharing data upon request, may not be sustainable for the long-term. For example, research has shown that ‘available by request’ does not work and furthermore that the availability of data declines rapidly with age.<sup>44</sup> Yet, data sharing is still happening and data curation efforts may help mitigate these error-prone approaches. Imker’s exploration of these “overlooked” methods will help data curators and librarians providing data services become better educated in the larger picture of scholarly data exchange.

## The Challenge of Providing Data Curation Services

In Part II, *Data Curation Services in Action*, we explore several examples of institutions already providing data curation services, review their service offerings, un-

derstand their technology infrastructure, and explore some of their challenging constraints, such as identifying appropriate cost-recovery models and rolling out promotion and marketing strategies that resonate with end users.

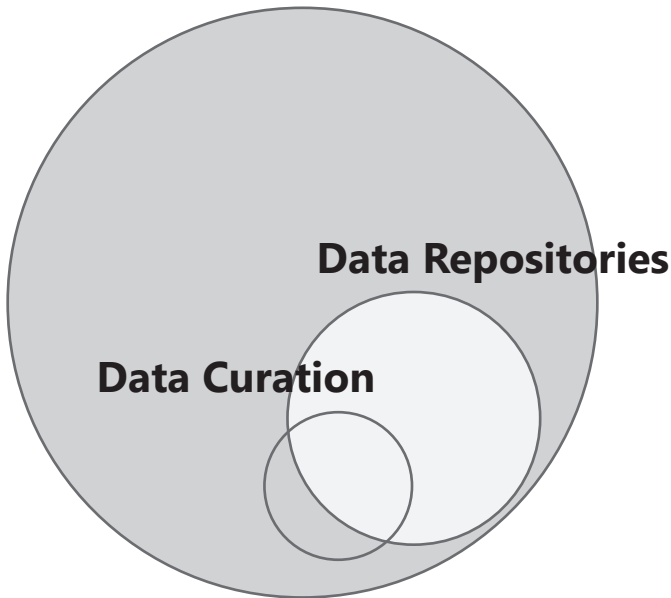
.....

In addition to the chapters described here, there are many practical examples to be found in this book's companion volume *Curating Research Data, Volume Two: A Handbook of Current Practice* which collects 30 practitioner case studies from institutional, disciplinary, and national data repositories in an eight-step workflow for data curation, from receiving to reuse.

.....

Putting data curation into context within the broader range of research data management services is essential as libraries shift toward progressively more responsible data stewardship roles at their institutions (see Figure Intro.1). For example, Witt describes the “information bottleneck” as a place where libraries can use data curation to help push valuable data sets beyond the laboratory and out to the broader research community.<sup>45</sup> Choudhury paints a rather bleak picture of the state of institutional repositories in 2008 and recommends data curation as a place of redemption for libraries in the larger scholarly communication landscape.<sup>46</sup> **In Chapter 6, authors Inna Kouper, Kathleen Fear, Mayu Ishida, Christine Kollen, and Sarah C. Williams address how far we have come with an empirical analysis of research data services provided by the Association of Research Libraries (ARL) in “Research Data Services Maturity in Academic Libraries.”** As the title suggests, the results of their study of current ARL service offerings are categorized by frequency into topographical levels and present a vocabulary for describing research data services (RDS). They find that basic services, such as data management plan consultations and data management workshops, were practiced in over 50% of their sample, while intermediate services, such as data deposit into repositories and data preservation, were only found in 15 percent to 50 percent of the group. Finally, the concept of data curation is found in less than 15 percent of the sample and labeled as an advanced service, which includes other services such as data and researcher IDs and data analysis. Their discussion of how these RDS concepts interrelate to one another provides an excellent snapshot at the evolving vernacular, if not actual nature, of our field. For example, the concept of data curation was still an emerging topic within the library science, archival, and information sciences disciplines just a few years ago and in fact very few academic libraries were successfully offering data curation services at all according to a study in 2011.<sup>47</sup> The RDS maturity model presents an opportunity to self-measure the actions our library takes in the broad arena of data services and allows us to strive to expand them to the next level.

## Research Data Services



### FIGURE INTRO.1

Data curation as a subset of research data services. Note that data curation services may support or overlap with local data repository services, or curation services may be provided for data that are deposited elsewhere, such as disciplinary repositories or non-accessible (dark) storage.

---

The next chapter in this volume provides an excellent case study in one academic library's ascendance from basic to advanced data services. **In Chapter 7, Jon Wheeler describes how academic library-run institutional repositories might be adapted to provide complementary platforms for data publication alongside disciplinary repositories in “Extending Data Curation Service Models for Academic Library and Institutional Repositories.”** Here the conflation between data sharing and data preservation come to a head. While academic researchers may deposit their data into disciplinary repositories to achieve one, then may not always be gaining the other. Wheeler presents data repository mirroring as one way for academic libraries to compliment successful disciplinary data repository efforts and goes on to provide several illustrative examples of “data mirroring” efforts underway with the University of New Mexico (UNM) Libraries. This example is unique by connecting an institutional repository to established disciplinary data repositories and collaborating their efforts. Disciplinary repositories such as Flybase, PLEXdb, and the Cambridge Structural

Database present the collective data outputs of a sub-topic in publicly accessible platforms designed to allow for widespread reuse of the data.<sup>48</sup> Within the context of disciplinary data repositories, several repository best practices for data curation emerge. For example, DataOne continues to educate the field by hosting workshops and publishing guides on research data management and software tools.<sup>49</sup> Their in-depth resources help researchers better prepare their data for eventual deposit into the DataOne connected archives.<sup>50</sup> Similarly detailed data curation instructions for oceanographic researchers are presented in the *Ocean Data Publication Cookbook*, which describes step-by-step instructions for curating disciplinary data from their field and applying digital object identifiers (DOIs) as a central component to the curation approach.<sup>51</sup>

Greater collaboration between the stakeholders of disciplinary and institutional data repositories would enhance our collective understanding of data curation best practices. In one area in particular there are several lessons to be learned: financial cost models for sustaining data repositories. Disciplinary data repositories have been grappling with how to maintain financial support beyond their initial start-up phase (often provided in the form of seed or grant funding) for decades.<sup>52</sup> For example, Ember and colleagues note the dichotomy between the long-term preservation costs of maintaining digital data, often indefinitely, with the periodic and uncertain grant support on which these repositories must rely.<sup>53</sup> Their white paper, resulting from a 2013 summit with representatives from twenty two disciplinary data repositories, evaluated several funding models and found both advantages and disadvantages. Their goals of meeting long-term sustainability, open access, and potential for equity by all depositors were not met by a single approach. For example, charging user fees to access data in the repository would limit open access, while depositor-incurred submission fees would lower equity for individual depositors not backed by generous grants or institutional open access funds. Only one approach (not currently in place in the US but found in other nations) appeared to provide a good balance: the infrastructure model. This was described as, “Funding agencies pay for archives directly as a necessary aspect of research infrastructure. The funding model is structured for long-term investment, rather than being tied to three-year grant cycles.”<sup>54</sup> **Chapter 8 draws from these cost models and many more in “Beyond Cost Recovery: Entrepreneurial Business Models for Data Curation in Academia,” in which Karl Nilsen reviews and compares the popular models for financing data curation efforts and reports on a new business model emerging at the University of Maryland Libraries.**

One potentially effective way to secure funding for your data repository may be to demonstrate positive use trends: both in data curation activities as well as reuse of the data your repository maintains. But the challenge here is determining how best to market and promote services to our intended audiences. **In Chapter 9, “Current Outreach and Marketing Practices for Research Data Repositories,” Katherine J. Gerwig from Metropolitan State University provides a mixed**

**methods approach to understanding the current data repository marketing and outreach strategies employed by over a dozen academic institutions.** Based on survey and interview results, Gerwig makes recommendations for those struggling to get the word out about their data curation services. For example, providing library liaisons, who are often embedded within their departmental cultures, with targeted messaging about the services in the form of presentation slides or an elevator speech was shown as one means of successful outreach activity. The lessons learned from current outreach efforts also demonstrates how libraries should reframe the data repository and curation efforts around the positive incentives for sharing data rather than the sharing requirements themselves: such as a means of advancing knowledge in their field or by facilitating reproduction and verification.

## Reuse: the Ultimate Goal of Data Curation?

Part III, Preparing Data for the Future, explores the outcomes of data curation efforts in numerous ways. If the ultimate goal of data curation is reuse, then how data are reused will inform the development of our services and best practices. But perhaps this is a thankless task? One illustrative quote comes from the introduction to a 2002 technical report, written by astronomer and Microsoft researcher Jim Gray, that aptly demonstrates the potentially uphill battle we face:

Once published, scientific data should remain available forever so that other scientists can reproduce the results and do new science with the data. Data may be used long after the project that gathered it ends. Later users will not implicitly know the details of how the data was gathered and prepared. To understand the data, those later users need the metadata: (1) how the instruments were designed and built; (2) when, where, and how the data was gathered; and (3) a careful description of the processing steps that led to the derived data products that are typically used for scientific data analysis. It's fine to say that scientists should record and preserve all this information, but it is far too laborious and expensive to document everything. The scientist wants to do science, not be a clerk. And besides, who cares? Most data is never looked at again anyway.<sup>55</sup>

The clarity and examples for types of “metadata” needed for successful data reuse in this example is impressive. Yet the sentiment that most data would not be looked at again does not hold up just over a decade later.

Instead, we are experiencing a dramatic shift in how data are reused, not only to “do new science,” but also because data reuse may increase a paper’s potential research impact, provide greater transparency to the results, and in some cases, can even make or break an individual’s career.<sup>56</sup> The research disciplines are often the driving force in the reproducibility (or replicability) movement using data sharing to build greater expectations for rerunning experiments, providing independent confirmations or validation of the research results, and more quickly identifying false findings.<sup>57</sup> Again, remembering that digital data are more easily shared, it is not surprising to ask researchers to provide the digital evidence of their findings for validation purposes. Some disciplines have embraced data transparency and provide portals and virtual hubs to share data and discuss results.<sup>58</sup> In one instance, national policy has embraced this idea of validation and Irish researchers are subject to external scrutiny when it comes to data presented in papers or captured in lab notebooks.<sup>59</sup>

Not everyone agrees that data transparency to the extreme is a positive trend. One 2016 editorial in *Nature* explains: “The progress of research demands transparency. But as scientists work to boost rigor, they risk making science more vulnerable to attacks. Awareness of tactics is paramount.”<sup>60</sup> They go on to provide 10 ways to “distinguish scrutiny from harassment.”<sup>61</sup> Another controversial take on data reuse issues erupted when the editor-in-chief of *The New England Journal of Medicine* (NEJM) published a sharply-worded editorial casting the role of data reuser as

...people who had nothing to do with the design and execution of the study but use another group’s data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as ‘research parasites.’<sup>62</sup>

A journalist from *Forbes* magazine drew an interesting comparable of the situation by suggesting, “In just four years, it seems, data science has devolved from the ‘sexiest job of the 21st century’ to a community of ‘research parasites,’” where the former linked to the widely cited *Harvard Business Review* report describing informatics-based jobs as exciting and lucrative career choices.<sup>63</sup> But the NEJM editorial, though sensational in some respects, does go on to make the point that researchers don’t want to be scooped, they don’t want to be proven wrong or taken out of context, and they are worried about not getting credit. Another researcher from a completely different field has a similar story. As co-author on a huge data sharing success story, the SnapShot Serengeti project hosted on the

community science driven platform Zooniverse, Kosmala describes some of the pressures faced by early career researchers to publish their results (in the form of traditional publications) and get scholarly credit for their work.<sup>64</sup> Data sharing, she argues, though admirable, removes overarching control over the data so that anyone else could use it, with your permission or not. On the other hand, when data are shared with conditions of co-authorship, the loss of control converts itself into an opportunity (even expectation) of collaboration. As data curators we must be keenly aware of these disincentives. Data sharing may be great for end users of data, but it can be not-so-great for the data creators. In addition to researcher fears, there are costs involved with data sharing in terms of time (and occasionally monetary investments), muddy ownership claims at stake, and well, data sharing can just be a “pain in the ass...”<sup>65</sup> In short, there is a lack of incentives for researchers to share: few carrots but many sticks.

Therefore, an additional role for data curators may be to understand and assist as much as possible in the ethical and appropriate reuse of data.

Library and information science professionals so often deal with the end-product in the scholarly communication pipeline, collecting the published finale of research: the papers, monographs, maps, and other well-formatted records of scholarship. Archives and special collections, on the other hand, cover a larger swath of the research process by also collecting the creation and evolution of a work in the form of an edited manuscript, unlabeled photos, and the order in which press clippings were arranged.<sup>66</sup> Research data curation may fall somewhere in between and be viewed as one way to bridge that gap of creation and final product by working with data creators to prepare their data for eventual publication, context and all. **In Chapter 10, “Open Exit: Reaching the End of the Data Lifecycle,” Andrea Ogier, Natsuko Nicholls, and Ryan Speer argue that data retention should be considered iteratively throughout the data life cycle and that knowledge gained from university records and information management, and library collection management can be applied to data curation efforts in order to assist with planned data obsolescence.** Rather than assume reuse potential for all data, our authors appropriately ask us to define better appraisal criteria to make critical selections for which data to retain and which data to dispose for reasons that incorporate the assessment of liability, risk, or resource cost over potential value.

But what happens once data have fallen into obsolescence? **Looking the opposite direction, Chapter 12 by Robert R. Downs and Robert S. Chen asks: when should data be resurrected? They describe the data curation actions that might be taken in order to protect data that are experiencing less than ideal conditions in “Curation of Scientific Data at Risk of Loss: Data Rescue and Dissemination.”** Their data rescue examples involve a data set that was originally housed in the National Biological Information Infrastructure (NBII) program of the United States Geological Survey (USGS). This repository is a



favorite among instructors of data information literacy due to its abrupt closure in response to federal budget cuts.<sup>67</sup> The digital archive was permanently taken offline in January 2012. Here our authors provide not only practical experiences from a data rescue effort but general advice on the benefits and challenges of these attempts. Their balanced recommendations to identify critical and timely documentation rather than strive for completeness are underscored by the relevant case study presented with the NBII dataset. Particularly notable are the intellectual property and ownership issues encountered with orphaned data as time passes, and their recommendation for data curators to apply metadata now, even at the most basic level, in order to help future curators pull out the details of the dataset in the possibly all-too-near future.

Finally, I'll close this introduction to Volume One with a focus on issues of worldwide access and discovery of data. This is an essential component of data curation and data discovery can be a key factor for prompting worldwide inclusivity in research. The 2005 NSB report projects that "Long-lived digital data collections are powerful catalysts for progress and for democratization of science and education."<sup>68</sup> Yet in 2015, Sorrono et al. argue that the inclusivity of data sharing is not well-discussed nor yet fully realized:

...a critical shift that is happening in both society and the environmental science community that makes data sharing not just good but ethically obligatory. This is a shift toward the ethical value of promoting inclusivity within and beyond science. An essential element of a truly inclusionary and democratic approach to science is to share data through publicly accessible data sets.<sup>69</sup>

Why? Because open data benefits science, enhances social and economic development, and, according to one Australian study, can even be significantly profitable.<sup>70</sup>

**In Chapter 11, "The Current State of Linked Data Repositories: A Comparative Analysis," Cynthia R. Hudson Vitale assesses the impact of the complexity of data sharing options available to researchers and observes that as a result data may be scattered across various institutional, disciplinary, or general repositories.** One possible solution is open and federated "meta-repositories" that search across the collective holdings of disparate data repositories. Lynch described this transition of data sharing practices as going from "journals [that] offer to accept it as 'supplementary materials' that accompany the article" to a future of repositories of machine-readable digital data that can be "data mined" for the generation of new knowledge.<sup>71</sup>

Hudson Vitale explores how this far end of the spectrum is emerging and compares thirteen linked data repositories, their underlying missions, and their technical approaches to federating data search and discovery using a website anal-

ysis across fifteen variables. The future of data reuse rests on the discoverability of data to potential reusers, and this chapter demonstrates that we have much to accomplish to make data repositories more interoperable.

## Conclusion

Digital data is ubiquitous and rapidly reshaping how scholarship progresses now and into the future. The abundant—and sometimes chaotic—flow of data worldwide enables a new form of collaborative exploration and discovery that minimizes international and interdisciplinary barriers connecting researchers with shared goals and accelerates the rate of scientific understanding. Just take a moment to consider the vast body of digital information housed in openly accessible data repositories across the world representing unique information products such as the mysterious and brief flashes of high-energy gamma-ray bursts originating from the far outer-reaches of our universe, the Alexandrian feat that is HathiTrust bringing together into a single corpus of searchable text everything from Shakespearean plays to song lyrics by The Beatles, the echoes of evolutionary history surfacing from the endless strings of human genetic DNA, and the daily snapshot of social norms and human values which can emerge from the deluge of human-machine interactions generated across the social web.<sup>72</sup> In 2003, Hey and Trefethen anticipated that “new types of digital libraries for scientific data with the same sort of management services as conventional digital libraries” would emerge in response to our changing world.<sup>73</sup> That time is now. These are extraordinary times for data curators and how we rise to the challenge of providing new services and respond to the shifting patterns of data sharing and data reuse has the potential to shape and define our profession into the future.

## Notes

1. Merriam-Webster’s Learner’s Dictionary, “Data,” accessed August 6, 2016, <http://www.merriam-webster.com/dictionary/data>.
2. Definition from footnote 1 on page 2 in the article by Claire C. Austin, Theodora Bloom, Sünje Dallmeier-Tiessen, Varsha K. Khodiyar, Fiona Murphy, Amy Nurnberger, Lisa Raymond, Martina Stockhause, Jonathan Tedds, Mary Vardigan, and Angus Whyte, “Key components of data publishing: Using current best practices to develop a reference model for data publishing,” *International Journal on Digital Libraries*, June 2016, doi:10.1007/s00799-016-0178-2.
3. See the Digital Curation Center (DCC). “DCC Curation Lifecycle Model,” accessed August 6, 2016, <http://www.dcc.ac.uk/resources/curation-lifecycle-model>; for the history and development of this model see Sarah Higgins, “The DCC Curation Lifecycle Model,” *International Journal of Digital Curation* 3, no. 1 (2008): 134–40, doi:10.2218/ijdc.v3i1.48, where data are defined on p137.

4. Ross Harvey, "Chapter 4. Defining Data," *Digital Curation: A How-To-Do-It Manual*, No. 025.06. (Chicago: Neal-Schuman Publishers, 2010), [http://www.alastore.ala.org/pdf/digital\\_curation.pdf](http://www.alastore.ala.org/pdf/digital_curation.pdf).
5. The US federal government, for example, defines research data in their OMB circular a-110 as "recorded factual material commonly accepted in the scientific community as necessary to validate research findings," see full notice at Office of Management and Budget, "CIRCULAR A-110," revised November 19, 1993, further amended September 20, 1999, [https://www.whitehouse.gov/omb/circulars\\_a110](https://www.whitehouse.gov/omb/circulars_a110).
6. See for example the PublicVR project, accessed August 6, 2016, <http://publicvr.org/index.html>, which provides virtual reality 3d environments for places such as the Grand Theater in the Roman city of Pompeii as it may have looked prior to the devastating volcanic eruption in 79AD.
7. See for example the eMotion lab at the University of Notre Dame that uses "advanced video capture equipment to track posture, gesture, and facial expression during a variety of experimental tasks" at the University of Notre Dame, "About the eMotion and eCognition Lab," accessed August 6, 2016, <http://www3.nd.edu/~emotecog/about.html>.
8. The 2015 report by McAfee Labs warns of the cyber security challenges that are abundant such as identity theft, data breaches, and national security risks in Intel Security Group McAfee Labs, "The Hidden Data Economy," October 15, 2015, <http://www.mcafee.com/us/resources/reports/rp-hidden-data-economy.pdf>; This Technology Watch report describes techniques to preserve large-scale transactional data derived from business and industry in Thomson, Sara Day, "Technology Watch Report 16: Preserving Transactional Data," Digital Preservation Coalition, May 2, 2016, doi:10.7207/twr16-02.
9. This quote is from page 2-1 of the OAIS Reference Model found in Consultative Committee for Space Data Systems, Audit and Certification of Trustworthy Digital Repositories, Recommended Practice, CCSDS 652.0-M-1, Magenta Book, Issue 1 Washington, DC: CCSDS Secretariat, September 2011, <http://public.ccsds.org/publications/archive/652x0m1.pdf>.
10. Footnote 2 on page 2 of Austin et. al. "Key components of data publishing: Using current best practices to develop a reference model for data publishing." Reference in the quote is to CASRAI, "Category:Research Data Domain," The CASRAI Dictionary, Last Modified August 18, 2015, [http://dictionary.casrai.org/Category:Research\\_Data\\_Domain](http://dictionary.casrai.org/Category:Research_Data_Domain); the RDA Data Foundations and Terminology working group has a growing dictionary of data related terms that is searchable at Research Data Alliance Data Foundation and Terminology Interest Group, "Term Definition Tool (TeD-T)," last modified March 1, 2016, [http://smw-rda.esc.rzg.mpg.de/index.php/Main\\_Page](http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page).
11. National Science Board, "NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century," Summer 2005, National Science Foundation, <http://www.nsf.gov/pubs/2005/nsb0540>, p1.
12. University of Illinois Urbana-Champaign School of Information Science, "Specialization in Data Curation," accessed August 4, 2016, [http://www.lis.illinois.edu/academics/programs/specializations/data\\_curation](http://www.lis.illinois.edu/academics/programs/specializations/data_curation).
13. Committee on Future Career Opportunities and Educational Requirements for Digital Curation; Board on Research Data and Information; Policy and Global Affairs; National Research Council, *Preparing the Workforce for Digital Curation* (Washington, DC: National Academies Press; April 22, 2015), [http://www.nap.edu/catalog.php?record\\_id=18590](http://www.nap.edu/catalog.php?record_id=18590).

14. For more in-depth coverage of this topic, read a systematic review of data sharing studies in academia. See: Fecher, Benedikt, Sascha Friesike, and Marcel Hebing, "What drives academic data sharing?," *PLoS One* 10, no. 2 (2015), doi:10.1371/journal.pone.0118053.
15. National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, *Information Technology and the Conduct of Research: The User's View* (Washington, DC: The National Academies Press, 1989), doi:10.17226/763, p1.
16. Gary King, "Ensuring the Data-Rich Future of the Social Sciences," *Science* 331(6018): 719–721 (2011), doi:10.1126/science.1197872.
17. An overview of these policies is found in Kathleen Shearer, "Comprehensive Brief on Research Data Management Policies," released April 2015, <http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157>.
18. The memo from the White House's Office of Science Technology Policy (OSTP) was released as John P. Holdren, "Increasing Access to the Results of Federally Funded Scientific Research," Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013, [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
19. Adapted from Inter-university Consortium for Political and Social Research (ICPSR), "Guidelines for OSTP Data Access Plan," accessed August 6, 2016, <http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/ostp.html>.
20. Jerry Sheehan, "Increasing Access to the Results of Federally Funded Science," The White House Blog, posted February 22, 2016, <https://www.whitehouse.gov/blog/2016/02/22/increasing-access-results-federally-funded-science>.
21. United States Government, "US Open Data Action Plan," May 9, 2014, [https://www.whitehouse.gov/sites/default/files/microsites/ostp/us\\_open\\_data\\_action\\_plan.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/us_open_data_action_plan.pdf).
22. Ford Foundation, "Ford Foundation expands Creative Commons licensing for all grant-funded projects," February 3, 2015, <https://www.fordfoundation.org/the-latest/news/ford-foundation-expands-creative-commons-licensing-for-all-grant-funded-projects>; Alfred P. Sloan Foundation, "Grant Application Guidelines," last modified January 6, 2014, [http://www.sloan.org/fileadmin/media/files/application\\_documents/proposal\\_guidelines\\_research\\_officer\\_grants.pdf](http://www.sloan.org/fileadmin/media/files/application_documents/proposal_guidelines_research_officer_grants.pdf); Bill & Melinda Gates Foundation, "Bill & Melinda Gates Foundation Open Access Policy," accessed August 6, 2016, <http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy>.
23. SPARC Open Data, "Research Funder Data Sharing Policies," accessed August 5, 2016, <http://sparcopen.org/our-work/research-data-sharing-policy-initiative/funder-policies>.
24. Institute of Medicine and National Academy of Sciences, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age* (Washington, DC: The National Academies Press, 2009), doi:10.17226/12615, 34.
25. Retraction Watch, "Archive for the 'data issues' Category," accessed August 6, 2016, <http://retractionwatch.com/category/by-reason-for-retraction/data-issues>.
26. Kathleen Fear, "Building Outreach on Assessment: Researcher Compliance with Journal Policies for Data Sharing," *Bulletin of the American Society for Information Science and Technology* 41, no. 6 (2015): 18–21, doi:10.1002/bult.2015.1720410609; Heather A. Piwowar and Wendy W. Chapman, "A Review of Journal Policies for Sharing Research Data," *Nature Precedings*, March 20, 2008, hdl:10101/npre.2008.1700.1; Linda Naughton and David Kernohan, "Making Sense of Journal Research Data Policies," *Insights*

- 29, no. 1 (2016), <http://doi.org/10.1629/uksg.284>.
27. The model is published in Paul Sturges, Marianne Bamkin, Jane H.S. Anders, Bill Hubbard, Azhar Hussain, and Melanie Heeley, "Research Data Sharing: Developing a Stakeholder-Driven Model for Journal Policies," *Journal of the Association for Information Science and Technology*, doi:10.1002/asi.23336.
  28. *Nature*, "Availability of Data, Material and Methods," accessed August 6, 2016, <http://www.nature.com/authors/policies/availability.html>; *PLOS One*, "Data Availability," accessed August 6, 2016, <http://journals.plos.org/plosone/s/data-availability>.
  29. Chelsey Coombs, "Neuroscience Paper Retracted After Colleagues Object to Data Publication," *Retraction Watch*, December 31, 2015, <http://retractionwatch.com/2015/12/31/neuroscience-paper-retracted-after-colleagues-object-to-data-publication>.
  30. Elsevier, "Elsevier and the Inter-University Consortium for Political and Social Research (ICPSR) Announce Data Linking," February 8, 2016, <http://www.prnewswire.com/news-releases/elsevier-and-the-inter-university-consortium-for-political-and-social-research-icpsr-announce-data-linking-568022141.html>; See the list of data repositories at Elsevier, "Supported Data Repositories," accessed August 6, 2016, <https://www.elsevier.com/?a=57755>.
  31. *Scientific Data* homepage, accessed August 6, 2016, <http://www.nature.com/sdata>; *Data in Brief* homepage, accessed August 6, 2016, <http://www.journals.elsevier.com/data-in-brief>; as reported in Tim Austin, "Towards a Digital Infrastructure for Engineering Materials Data," *Materials Discovery* (2016), doi:10.1016/j.md.2015.12.003, 2.
  32. Leonardo Candela, Donatella Castelli, Paolo Manghi, and Alice Tani, "Data Journals: A Survey," *Journal of the Association for Information Science and Technology* 66, no. 9 (2015): 1747–1762, doi: 10.1002/asi.23358.
  33. *Ibid*, 1756.
  34. *Scientific Data*, "Recommended Data Repositories," accessed July 18, 2016, <http://www.nature.com/sdata/policies/repositories>.
  35. The declaration signifies that each country will "Work towards the establishment of access regimes for digital research data from public funding" and with shared objectives and principles. Available as Organisation for Economic Co-operation and Development, "Declaration on Access to Research Data from Public Funding," January 30, 2004, <http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157>.
  36. The UK funding council policies are each summarized and linked to from the Digital Curation Center, "Funders' Data Policies," accessed August 6, 2016, <http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies>; the Wellcome Trust, "Policy on data management and sharing," accessed August 6, 2016, <https://wellcome.ac.uk/funding/managing-grant/policy-data-management-and-sharing>; Research Councils UK, "RCUK Common Principles on Data Policy," published April 2011, <http://www.rcuk.ac.uk/research/datapolicy>.
  37. European Commission, "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020", version 3.0," July 26, 2016, [http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-pilot-guide\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf).
  38. Kathleen Shearer, "Comprehensive Brief on Research Data Management Policies." In 2015 Canada also released a federal policy on the open access to publications resulting from federal funds from its three primary funding agencies (see Government of Canada,

- “Tri-Agency Open Access Policy on Publications,” February 27, 2015, <http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1>, yet this requirement only applies to research articles, not data.
39. Portage network homepage, accessed August 6, 2016, <https://portagenetwork.ca>.
  40. JISC-funded Research Data Management Shared Service Project, accessed August 4, 2016, <https://www.jisc.ac.uk/rd/projects/research-data-shared-service>; Data Curation Network Project homepage, accessed August 4, 2016, <https://sites.google.com/site/data-curationnetwork>.
  41. For example, findings from reviewing a sample of 182 Data Management Plans of successful National Science Foundation grant proposals showed this to be the case for 74% of the sample in Carolyn Bishoff and Lisa R. Johnston, “Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University,” *Journal of Librarianship and Scholarly Communication* 3, no. 2 (2015). doi:10.7710/2162-3309.1231.
  42. Caitlin Rivers, “‘Send Me Your Data—PDF is Fine,’ Said No One Ever (How to Share Your Data Effectively),” April 8, 2013, <http://www.caitlinrivers.com/blog/send-me-your-data-pdf-is-fine-said-no-one-ever-how-to-share-your-data-effectively>.
  43. Carlos Santos, Judith Blake, and David J. States, “Supplementary Data Need to be Kept in Public Repositories,” *Nature* 438, no. 7069 (2005): 738–738, doi: 10.1038/438738a.
  44. Caroline J. Savage, and Andrew J. Vickers, “Empirical Study of Data Sharing by Authors Publishing in PLoS Journals,” *PLoS One* 4, no. 9 (2009): e7078, doi:10.1371/journal.pone.0007078; Timothy H. Vines, Arianne YK Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison, “The Availability of Research Data Declines Rapidly with Article Age,” *Current Biology* 24, no. 1 (2014): 94–97, doi:10.1016/j.cub.2013.11.014.
  45. Michael Witt, “Institutional Repositories and Research Data Curation in a Distributed Environment,” *Library Trends* 57, no. 2 (2008): 191–201, doi:10.1353/lib.0.0029.
  46. G. Sayeed Choudhury, “Case Study in Data Curation at Johns Hopkins University,” *Library Trends* 57, no. 2 (2008): 211–220, doi:10.1353/lib.0.0028.
  47. Carol Tenopir, Ben Birch, and Suzie Allard, *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*, An ACRL White Paper, Association of College and Research Libraries, a division of the American Library Association, 2012, [http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir\\_Birch\\_Allard.pdf](http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf).
  48. Further examples of disciplinary repositories are found in re3data.org homepage, accessed August 6, 2016, <http://www.re3data.org>.
  49. DataOne, “Best Practices,” accessed August 5, 2016, <http://www.dataone.org/best-practices>; DataOne, “Software Tools Catalog,” accessed August 5, 2016, [https://www.dataone.org/software\\_tools\\_catalog](https://www.dataone.org/software_tools_catalog).
  50. DataOne, “ESA 2011: How to Manage Ecological Data for Effective Use and Re-use,” August 7, 2011, <http://www.dataone.org/esa-2011-how-manage-ecological-data-effective-use-and-re-use>.
  51. Raymond Leadbetter, A. L., Chandler, C., Pikula, L., Pissierssens, P., Urban, E., *Ocean Data Publication Cookbook* (Paris: UNESCO, 2013), <http://www.iode.org/mg64>; For further context see the slides by Lisa Raymond, “Publishing and Citing Ocean Data,” OneNOAA Science Seminar, National Oceanographic Data Center, May 22, 2013,

- [http://www.nodc.noaa.gov/seminars/2013/support/Lisa\\_Raymond\\_OneNOAASeminar\\_slides.pdf](http://www.nodc.noaa.gov/seminars/2013/support/Lisa_Raymond_OneNOAASeminar_slides.pdf).
52. Jared Lyle, George Alter and Mary Vardigan, “‘The Price of Keeping Knowledge’ Workshop: ICPSR Position Paper,” (2013), [http://www.knowledge-ex-change.info/Admin/Public/DWSDownload.aspx?File=%2FFiles%2FFiler%2Fdownloads%2FPrimary+Research+Data%2FWorkshop+Price+of+Keeping+Knowledge%2FJared+Lyle+ICPSR\\_Position+Paper\\_Price+workshop\\_public.pdf](http://www.knowledge-ex-change.info/Admin/Public/DWSDownload.aspx?File=%2FFiles%2FFiler%2Fdownloads%2FPrimary+Research+Data%2FWorkshop+Price+of+Keeping+Knowledge%2FJared+Lyle+ICPSR_Position+Paper_Price+workshop_public.pdf).
  53. Carol Ember, Robert Hanisch, George Alter, Helen Berman, Margaret Hedstrom, and Mary Vardigan. “Sustaining Domain Repositories for Digital Data: A White Paper,” December 11, 2013, 10–11, [http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper\\_ICPSR\\_SDRDD\\_121113.pdf](http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf).
  54. *Ibid.*, 10.
  55. Jim Gray, Alexander S. Szalay, Ani R. Thakar, Christopher Stoughton, and Jan vandenBerg, “Online Scientific Data Curation, Publication, and Archiving,” submitted August 7, 2002, <http://arxiv.org/abs/cs.DL/0208012>.
  56. According to a 2007 study, openly sharing data was linked higher citation rates for the publications associated with that data. See Heather A. Piwowar, Roger S. Day, and Douglas B. Fridms, “Sharing Detailed Research Data is Associated with Increased Citation Rate,” *PLoS One* 2, no. 3 (2007): e308, doi:10.1371/journal.pone.0000308; Cases of unreplicable or faulty data have been the subject of several studies, such as the Reproducibility Studies by the Center for Open Science in the fields of psychology, (Alexander A. Aarts, Christopher J. Anderson, Joanna Anderson, Marcel A.L.M van Assen, Peter R. Attridge, Angela S. Attwood, Jordan Axt, et al., 2016, “Reproducibility Project: Psychology,” Open Science Framework, July 23, <https://osf.io/EZcUj/>); and cancer biology (Timothy M. Errington, Fraser E. Tan, Joelle Lomax, Nicole Perfito, Elizabeth Iorns, William Gunn, Brian A. Nosek, et al., 2016, “Reproducibility Project: Cancer Biology,” Open Science Framework, July 22. <https://osf.io/e81xl/>). In addition, the high profile case of scientists Dong-Pyou Han in an HIV-data falsification charge actually led to jail time and \$7.2 million in fines according to the report Sara Reardon, “US Vaccine Researcher Sentenced to Prison for Fraud,” *Nature News*, July 1, 2015, <http://www.nature.com/news/us-vaccine-researcher-sentenced-to-prison-for-fraud-1.17660>.
  57. Victoria Sodden provides entertaining slide presentation on “A Brief History of the Reproducibility Movement,” December 10, 2012, <http://hdl.handle.net/10022/AC:P:15396>; Prasad Patil, Roger D. Peng, Jeffrey Leek, “A Statistical Definition for Reproducibility and Replicability,” *BioRxiv*, July 29, 2016, doi:10.1101/066803.
  58. Disciplinary repositories such as the iPlant Collaborative (homepage, accessed August 6, 2016, <http://www.iplantcollaborative.org>), nanoHUB.org (homepage, accessed August 6, 2016, <https://nanohub.org>), EarthCube (homepage, accessed August 6, 2016, <http://earthcube.org>), and CUAHSI (Hydrologic Information System homepage, accessed August 6, 2016, <http://his.cuahsi.org>) represent the collective outputs of the discipline to allow for widespread reuse of the data.
  59. Richard Van Noorden, “Irish University Labs Face External Audits,” *Nature News*, June 17, 2014, <http://www.nature.com/news/irish-university-labs-face-external-audits-1.15422>.
  60. Stephan Lewandowsky and Dorothy Bishop, “Research Integrity: Don’t Let Transparency Damage Science,” *Nature*, January 25, 2016, <http://www.nature.com/news/research-integrity-don-t-let-transparency-damage-science-1.19219>.



61. Ibid.
62. Dan L. Longo, and Jeffrey M. Drazen, "Data Sharing," *New England Journal of Medicine* 374, no. 3 (2016): 276–277, doi: 10.1056/NEJMe1516564.
63. David Shaywitz, "Data Scientists = Research Parasites?," *Forbes*, January 21, 2016, <http://www.forbes.com/sites/davidshaywitz/2016/01/21/data-scientists-research-parasites/#3ddef3453d1c>; Thomas H. Davenport and D.J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, October 2012, <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>.
64. Margaret Kosmala, "Open Data, Authorship, and the Early Career Scientist," *Ecology Bits*, posted June 15, 2016, <http://ecologybits.com/index.php/2016/06/15/open-data-authorship-and-the-early-career-scientist/>; Snapshot Serengeti dataset available as Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer, "Snapshot Serengeti, High-Frequency Annotated Camera Trap Images of 40 Mammalian Species in an African Savanna," *Dryad Digital Repository*, <http://dx.doi.org/10.5061/dryad.5pt92> and the paper describing the data available as Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer, "Snapshot Serengeti, High-Frequency Annotated Camera Trap Images of 40 Mammalian Species in an African Savanna," *Scientific Data* 2 (2015), doi:10.1038/sdata.2015.26.
65. Terry McGlynn, "I Own My Data, Until I Don't," *Small Pond Science*, March 3, 2014, <http://smallpondscience.com/2014/03/03/i-own-my-data-until-i-dont>; Emilio M. Bruna, "The Opportunity Cost of My #OpenScience was 36 Hours + \$690," The Bruma Lab, September 4, 2014, <http://brunalab.org/blog/2014/09/04/the-opportunity-cost-of-my-openscience-was-35-hours-690>.
66. The archival community has dealt with curation issues in the print and analog for centuries and the lessons learned translate well into the digital realm but are often overlooked by developers of new data curation services in academic and disciplinary settings according to Helen R. Tibbo, and Christopher A. Lee, "Closing the Digital Curation Gap: A Grounded Framework for Providing Guidance and Education in Digital Curation," Archiving Conference, vol. 2012, no. 1, pp. 57–62, *Society for Imaging Science and Technology*, 2012, <http://www.ils.unc.edu/calcee/p57-tibbo.pdf>. Some example archival workflows that translate well to data curation include Julianna Barrera-Gomez and Ricky Erway, *Walk This Way: Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-House* (Dublin, OH: OCLC Online Computer Library Center, 2013), <http://www.oclc.org/content/dam/research/publications/library/2013/2013-02.pdf> and the AIMS Work Group, "AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship," January 2012, [http://dcs.library.virginia.edu/files/2013/02/AIMS\\_final.pdf](http://dcs.library.virginia.edu/files/2013/02/AIMS_final.pdf).
67. US Geological Survey, "NBII to Be Taken Offline Permanently in January," *USGS Access Newsletter* 14, no. 3 (Fall 2011), [https://www2.usgs.gov/core\\_science\\_systems/Access/p1111-1.html](https://www2.usgs.gov/core_science_systems/Access/p1111-1.html).
68. National Science Board, "NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century," <https://www.nsf.gov/pubs/2005/nsb0540/>.
69. Patricia A. Soranno, Kendra S. Cheruvelil, Kevin C. Elliott, and Georgina M. Montgomery, "It's Good to Share: Why Environmental Scientists' Ethics are Out of Date," *BioScience* 65, no. 1 (2015): 69–73, doi: 10.1093/biosci/biu169.
70. Australian National Data Service, "Open Research Data," November 2014, <http://www>.



ands.org.au/working-with-data/articulating-the-value-of-open-data/open-research-data-report.

71. Clifford Lynch, “The Shape of the Scientific Article in the Developing Cyberinfrastructure,” *CTWatch Quarterly* 3, no. 3 (2007), <http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/index.html>.
72. Real-time observational data of the quickly dimming objects known as gamma-ray bursts (GRBs) are available to researchers through the Goddard Space Flight Center, “GCN: The Gamma-ray Coordinates Network (TAN: Transient Astronomy Network),” accessed August 6, 2016, <http://gcn.gsfc.nasa.gov> and public download access to GRB recordings that predate the SWIFT satellite mission launched in 2003 are also available Goddard Space Flight Center, “The Gamma Ray Burst Catalog,” accessed August 6, 2016, <http://heasarc.gsfc.nasa.gov/grbcatalog/grbcatalog.html>; Hathitrust is a searchable database of millions of digitized text and available at Hathitrust homepage, accessed August 6, 2016, <http://babel.hathitrust.org>; Public access to download the human genome and tools to analyze and compare DNA are available at NCBI, “Human Genome Resources,” accessed August 6, 2016, <http://www.ncbi.nlm.nih.gov/genome/guide/human>; Big data generated by human-computer interaction can be derived from many social web services, though some do not release their data to the public (e.g., Amazon, Facebook). Sources of public data are available via APIs that contain real-time, and sometimes historical, information. For example Twitter interaction data can be found at the Gnip homepage, accessed August 6, 2016, <https://gnip.com>, and in 2016 Yahoo released a News Feed dataset of 110 billion interactions of anonymized users interactions with their home page and news sites as Yahoo, “R10—Yahoo News Feed dataset, version 1.0 (1.5TB),” accessed August 6, 2016, <http://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75>.
73. Anthony J.G. Hey, and Anne E. Trefethen, “The Data Deluge: An E-Science Perspective,” *Grid Computing: Making the Global Infrastructure a Reality*, (Chichester: Wiley, 2003), 809–24, <http://eprints.soton.ac.uk/id/eprint/257648>.

## Bibliography

- Aarts, Alexander A., Christopher J. Anderson, Joanna Anderson, Marcel A.L.M van Assen, Peter R. Attridge, Angela S. Attwood, Jordan Axt, et al. 2016. “Reproducibility Project: Psychology.” Open Science Framework. July 23. [osf.io/ezcuj](https://osf.io/ezcuj).
- AIMS Work Group. “AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.” January 2012. [http://dcs.library.virginia.edu/files/2013/02/AIMS\\_final.pdf](http://dcs.library.virginia.edu/files/2013/02/AIMS_final.pdf).
- Alfred P. Sloan Foundation. “Grant Application Guidelines.” Last modified January 6, 2014. [http://www.sloan.org/fileadmin/media/files/application\\_documents/proposal\\_guidelines\\_research\\_officer\\_grants.pdf](http://www.sloan.org/fileadmin/media/files/application_documents/proposal_guidelines_research_officer_grants.pdf).
- Austin, Claire C., Theodora Bloom, Sünje Dallmeier-Tiessen, Varsha K. Khodiyar, Fiona Murphy, Amy Nurnberger, Lisa Raymond, Martina Stockhause, Jonathan Tedds, Mary Vardigan, and Angus Whyte. “Key components of data publishing: Using current best practices to develop a reference model for data publishing.” *International Journal on Digital Libraries*, 20 June 2016. doi:10.1007/s00799-016-0178-2.

- Austin, Tim. "Towards a Digital Infrastructure for Engineering Materials Data." *Materials Discovery* (2016). doi:10.1016/j.md.2015.12.003.
- Australian National Data Service. "Open Research Data." November 2014. <http://www.ands.org.au/working-with-data/articulating-the-value-of-open-data/open-research-data-report>.
- Barrera-Gomez, Julianna, and Ricky Erway. *Walk This Way: Detailed Steps for Transferring Born-Digital Content from Media You Can Read In-House*. Dublin, OH: OCLC Online Computer Library Center, Inc., 2013. <http://www.oclc.org/content/dam/research/publications/library/2013/2013-02.pdf>.
- Bill & Melinda Gates Foundation. "Bill & Melinda Gates Foundation Open Access Policy." Accessed August 6, 2016. <http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy>.
- Bishoff, Carolyn, and Lisa R. Johnston. "Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University." *Journal of Librarianship and Scholarly Communication* 3, no. 2 (2015). doi:10.7710/2162-3309.1231.
- Bruna, Emilio M. "The Opportunity Cost of My #OpenScience was 36 Hours + \$690." *The Bruma Lab*. September 4, 2014. <http://brunalab.org/blog/2014/09/04/the-opportunity-cost-of-my-openscience-was-35-hours-690/>.
- Candela, Leonardo, Donatella Castelli, Paolo Manghi, and Alice Tani. "Data Journals: A Survey." *Journal of the Association for Information Science and Technology* 66, no. 9 (2015): 1747-1762. doi: 10.1002/asi.23358.
- CASRAI. "Category:Research Data Domain." The CASRAI Dictionary. Last Modified August 18, 2015. [http://dictionary.casrai.org/Category:Research\\_Data\\_Domain](http://dictionary.casrai.org/Category:Research_Data_Domain).
- Choudhury, G. Sayeed. "Case Study in Data Curation at Johns Hopkins University." *Library Trends* 57, no. 2 (2008): 211-220. doi: 10.1353/lib.0.0028.
- Committee on Future Career Opportunities and Educational Requirements for Digital Curation; Board on Research Data and Information; Policy and Global Affairs; National Research Council. *Preparing the Workforce for Digital Curation*. Washington, DC: National Academies Press; April 22, 2015. [http://www.nap.edu/catalog.php?record\\_id=18590](http://www.nap.edu/catalog.php?record_id=18590).
- Consultative Committee for Space Data Systems. Audit and Certification of Trustworthy Digital Repositories. Recommended Practice, CCSDS 652.0-M-1, Magenta Book, Issue 1. Washington, DC: CCSDS Secretariat, September 2011. <http://public.ccsds.org/publications/archive/652x0m1.pdf>.
- Coombs, Chelsey. "Neuroscience Paper Retracted After Colleagues Object to Data Publication." *Retraction Watch*. December 31, 2015. <http://retractionwatch.com/2015/12/31/neuroscience-paper-retracted-after-colleagues-object-to-data-publication/>.
- CUAHSI Hydrologic Information System homepage. Accessed August 6, 2016. <http://his.cuahsi.org/>.
- Data Curation Network Project homepage. Accessed August 4, 2016. <https://sites.google.com/site/datacurationnetwork/>.
- Data in Brief homepage. Accessed August 6, 2016. <http://www.journals.elsevier.com/data-in-brief>.
- DataOne. "Best Practices." Accessed August 5, 2016. <http://www.dataone.org/best-practices>.

- DataOne. "ESA 2011: How to Manage Ecological Data for Effective Use and Re-use." August 7, 2011. <http://www.dataone.org/esa-2011-how-manage-ecological-data-effective-use-and-re-use>.
- DataOne. "Software Tools Catalog." Accessed August 5, 2016. [https://www.dataone.org/software\\_tools\\_catalog](https://www.dataone.org/software_tools_catalog).
- Davenport, Thomas H., D.J. Patil. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review*. October 2012. <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>.
- Digital Curation Center. "Funders' Data Policies." Accessed August 6, 2016. <http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies>.
- Digital Curation Center (DCC). "DCC Curation Lifecycle Model." Accessed August 6, 2016. <http://www.dcc.ac.uk/resources/curation-lifecycle-model>.
- EarthCube homepage. Accessed August 6, 2016. <http://earthcube.org/>.
- Elsevier. "Elsevier and the Inter-University Consortium for Political and Social Research (ICPSR) Announce Data Linking." February 8, 2016. <http://www.prnewswire.com/news-releases/elsevier-and-the-inter-university-consortium-for-political-and-social-research-icpsr-announce-data-linking-568022141.html>.
- . "Supported Data Repositories." Accessed August 6, 2016. <https://www.elsevier.com/?a=57755>.
- Ember, Carol, Robert Hanisch, George Alter, Helen Berman, Margaret Hedstrom, and Mary Vardigan. "Sustaining Domain Repositories for Digital Data: A White Paper." December 11, 2013, 10–11. [http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper\\_ICPSR\\_SDRDD\\_121113.pdf](http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf).
- Errington, Timothy M, Fraser E. Tan, Joelle Lomax, Nicole Perfito, Elizabeth Iorns, William Gunn, Brian A. Nosek, et al. 2016. "Reproducibility Project: Cancer Biology." Open Science Framework. July 22. [osf.io/e81xl](https://osf.io/e81xl).
- European Commission. "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Version 3.0." July 26, 2016. [http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-pilot-guide\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf).
- Fear, Kathleen. "Building Outreach on Assessment: Researcher Compliance with Journal Policies for Data Sharing." *Bulletin of the American Society for Information Science and Technology* 41, no. 6 (2015): 18-21. doi:10.1002/bult.2015.1720410609.
- Fecher, Benedikt, Sascha Friesike, and Marcel Hebing. "What Drives Academic Data Sharing?" *PLoS One* 10, no. 2 (2015): doi:10.1371/journal.pone.0118053.
- Ford Foundation. "Ford Foundation expands Creative Commons licensing for all grant-funded projects." February 3, 2015. <https://www.fordfoundation.org/the-latest/news/ford-foundation-expands-creative-commons-licensing-for-all-grant-funded-projects/>.
- Gnip homepage. Accessed August 6, 2016. <https://gnip.com/>.
- Goddard Space Flight Center. "GCN: The Gamma-ray Coordinates Network (TAN: Transient Astronomy Network)." Accessed August 6, 2016. <http://gcn.gsfc.nasa.gov>.
- Goddard Space Flight Center. "The Gamma Ray Burst Catalog." Accessed August 6, 2016. <http://heasarc.gsfc.nasa.gov/grbcatalog/grbcatalog.html>.
- Government of Canada. "Tri-Agency Open Access Policy on Publications." February 27, 2015. <http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1>.
- Gray, Jim, Alexander S. Szalay, Ani R. Thakar, Christopher Stoughton, and Jan vandenBerg. "Online Scientific Data Curation, Publication, and Archiving." Submitted August 7, 2002. <http://arxiv.org/abs/cs.DL/0208012>.

- Harvey, Ross. "Chapter 4. Defining Data." *Digital Curation: A How-To-Do-It Manual*. No. 025.06. Chicago: Neal-Schuman Publishers, 2010.
- HathiTrust homepage. Accessed August 6, 2016. <http://babel.hathitrust.org>.
- Hey, Anthony J.G., and Anne E. Trefethen. "The Data Deluge: An E-Science Perspective." In *Grid Computing: Making the Global Infrastructure a Reality*, edited by F. Berman, G. Fox, A. J.G. Hey, 809–24. Chichester: Wiley 2003. <http://eprints.soton.ac.uk/id/eprint/257648>.
- Higgins, Sarah. "The DCC Curation Lifecycle Model." *International Journal of Digital Curation* 3, no. 1 (2008): 134–40. doi:10.2218/ijdc.v3i1.48, p137.
- Holdren, John P. "Increasing Access to the Results of Federally Funded Scientific Research." Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013. [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
- Institute of Medicine and National Academy of Sciences. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington, DC: The National Academies Press, 2009. doi:10.17226/12615, 34.
- Intel Security Group McAfee Labs. "The Hidden Data Economy." October 15, 2015. <http://www.mcafee.com/us/resources/reports/rp-hidden-data-economy.pdf>.
- Inter-university Consortium for Political and Social Research (ICPSR). "Guidelines for OSTP Data Access Plan." Accessed August 6, 2016. <http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/ostp.html>.
- iPlant Collaborative homepage. Accessed August 6, 2016. <http://www.iplantcollaborative.org>.
- King, Gary. 2011. Ensuring the Data-rich Future of the Social Sciences. *Science* 331(6018): 719–721. doi:10.1126/science.1197872.
- Kosmala, Margaret. "Open Data, Authorship, and the Early Career Scientist." *Ecology Bits*, posted June 15, 2016. <http://ecologybits.com/index.php/2016/06/15/open-data-authorship-and-the-early-career-scientist>.
- Leadbetter, A., Raymond, L., Chandler, C., Pikula, L., Pissierssens, P., Urban, E. *Ocean Data Publication Cookbook*. (Paris: UNESCO, 2013.) <http://www.iode.org/mg64>.
- Lewandowsky, Stephan and Dorothy Bishop. "Research Integrity: Don't Let Transparency Damage Science." *Nature*. January 25, 2016. <http://www.nature.com/news/research-integrity-don-t-let-transparency-damage-science-1.19219>.
- Longo, Dan L. and Jeffrey M. Drazen. "Data Sharing." *New England Journal of Medicine* 374, no. 3 (2016): 276-277. doi:10.1056/NEJMe1516564.
- Lyle, Jared, George Alter, and Mary Vardigan. "The Price of Keeping Knowledge Workshop: ICPSR Position Paper." (2013) [http://www.knowledge-ex-change.info/Admin/Public/DWSDownload.aspx?File=%2FFiles%2FFiler%2Fdownloads%2FPrimary+Research+Data%2FWorkshop+Price+of+Keeping+Knowledge%2FJared+Lyle+ICPSR+Position+Paper+Price+workshop\\_public.pdf](http://www.knowledge-ex-change.info/Admin/Public/DWSDownload.aspx?File=%2FFiles%2FFiler%2Fdownloads%2FPrimary+Research+Data%2FWorkshop+Price+of+Keeping+Knowledge%2FJared+Lyle+ICPSR+Position+Paper+Price+workshop_public.pdf).
- Lynch, Clifford. "The Shape of the Scientific Article in the Developing Cyberinfrastructure." *CTWatch Quarterly* 3, no. 3 (2007). <http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/index.html>.
- McGlynn, Terry. "I Own My Data, Until I Don't." *Small Pond Science*. March 3, 2014. <http://smallpondscience.com/2014/03/03/i-own-my-data-until-i-dont/>.

- Merriam-Webster's Learner's Dictionary. "Data." Web version. Accessed August 6, 2016. <http://www.merriam-webster.com/dictionary/data>.
- nanoHUB.org homepage. Accessed August 6, 2016. <https://nanohub.org/>.
- National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. *Information Technology and the Conduct of Research: The User's View*. Washington, DC: The National Academies Press, 1989. doi:10.17226/763.
- National Science Board. "NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century." Summer 2005. National Science Foundation. <http://www.nsf.gov/pubs/2005/nsb0540>.
- Nature*. "Availability of Data, Material and Methods." Accessed August 6, 2016. <http://www.nature.com/authors/policies/availability.html>.
- Naughton, Linda and David Kernohan. "Making Sense of Journal Research Data Policies." *Insights* 29, no. 1 (2016). doi: <http://doi.org/10.1629/uksg.284>.
- NCBI. "Human Genome Resources." Accessed August 6, 2016. <http://www.ncbi.nlm.nih.gov/genome/guide/human>.
- Office of Management and Budget. "CIRCULAR A-110." Revised November 19, 1993 as further amended September 20, 1999. [https://www.whitehouse.gov/omb/circulars\\_a110\\_OMB\\_circular\\_a-110](https://www.whitehouse.gov/omb/circulars_a110_OMB_circular_a-110).
- Organisation for Economic Co-operation and Development. "Declaration on Access to Research Data from Public Funding." January 30, 2004. <http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157>.
- Patil, Prasad, Roger D. Peng, and Jeffrey Leek. "A Statistical Definition for Reproducibility and Replicability." *BioRxiv*. July 29, 2016. doi:10.1101/066803.
- Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. "Sharing Detailed Research Data is Associated with Increased Citation Rate." *PLoS One* 2, no. 3 (2007): e308. doi:10.1371/journal.pone.0000308.
- Piwowar, Heather A. and Wendy W. Chapman. "A Review of Journal Policies for Sharing Research Data." *Nature Precedings*. March 20, 2008. hdl:10101/npre.2008.1700.1. *PLOS One*. "Data Availability." Accessed August 6, 2016. <http://journals.plos.org/plosone/s/data-availability>.
- Portage network homepage. Accessed August 6, 2016. <https://portagenetwork.ca/>.
- PublicVR project homepage. Accessed August 6, 2016. <http://publicvr.org/index.html>.
- Raymond, Lisa. "Publishing and Citing Ocean Data." One NOAA Science Seminar, National Oceanographic Data Center. May 22, 2013. [http://www.nodc.noaa.gov/seminars/2013/support/Lisa\\_Raymond\\_OneNOAASeminar\\_slides.pdf](http://www.nodc.noaa.gov/seminars/2013/support/Lisa_Raymond_OneNOAASeminar_slides.pdf).
- re3data.org homepage. Accessed August 6, 2016. <http://www.re3data.org/>.
- Reardon, Sara. "US Vaccine Researcher Sentenced to Prison for Fraud." *Nature News*, July 1, 2015. <http://www.nature.com/news/us-vaccine-researcher-sentenced-to-prison-for-fraud-1.17660>.
- Research Councils UK. "RCUK Common Principles on Data Policy." April 2011. <http://www.rcuk.ac.uk/research/datapolicy/>.
- Research Data Alliance Data Foundation and Terminology Interest Group. "Term Definition Tool (TeD-T)." Last modified March 1, 2016. [http://smw-rda.esc.rzg.mpg.de/index.php/Main\\_Page](http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page).
- Research Data Management Shared Service Project homepage. Accessed August 4, 2016. <https://www.jisc.ac.uk/rd/projects/research-data-shared-service>.

- Retraction Watch. "Archive for the 'Data Issues' Category." Accessed August 6, 2016. <http://retractionwatch.com/category/by-reason-for-retraction/data-issues/>.
- Rivers, Caitlin. "'Send Me Your Data - PDF is Fine,' Said No One Ever (How to Share Your Data Effectively)." April 8, 2013. <http://www.caitlinrivers.com/blog/send-me-your-data-pdf-is-fine-said-no-one-ever-how-to-share-your-data-effectively>.
- Santos, Carlos, Judith Blake and David J. States. "Supplementary Data Need to be Kept in Public Repositories." *Nature* 438, no. 7069 (2005): 738-738. doi: 10.1038/438738a.
- Savage, Caroline J. and Andrew J. Vickers. "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals." *PLoS One* 4, no. 9 (2009): e7078. doi:10.1371/journal.pone.0007078.
- Scientific Data* homepage. Accessed August 6, 2016. <http://www.nature.com/sdata>.
- Scientific Data*. "Recommended Data Repositories." Accessed July 18, 2016. <http://www.nature.com/sdata/policies/repositories>.
- Shaywitz, David. "Data Scientists = Research Parasites?" *Forbes*, January 21, 2016. <http://www.forbes.com/sites/davidshaywitz/2016/01/21/data-scientists-research-parasites/#3ddef3453d1c>.
- Shearer, Kathleen. "Comprehensive Brief on Research Data Management Policies." Released April 2015. <http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157>.
- Sheehan, Jerry. "Increasing Access to the Results of Federally Funded Science." The White House Blog. February 22, 2016. <https://www.whitehouse.gov/blog/2016/02/22/increasing-access-results-federally-funded-science>.
- Sodden, Victoria. "A Brief History of the Reproducibility Movement." December 10, 2012. <http://hdl.handle.net/10022/AC:P:15396>.
- Soranno, Patricia A., Kendra S. Cheruvilil, Kevin C. Elliott, and Georgina M. Montgomery. "It's Good to Share: Why Environmental Scientists' Ethics are Out of Date." *BioScience* 65, no. 1 (2015): 69-73. doi: 10.1093/biosci/biu169.
- SPARC Open Data. "Research Funder Data Sharing Policies." Accessed August 5, 2016. <http://sparcopen.org/our-work/research-data-sharing-policy-initiative/funder-policies/>.
- Sturges, Paul, Marianne Bamkin, Jane H.S. Anders, Bill Hubbard, Azhar Hussain and Melanie Heeley. "Research Data Sharing: Developing a Stakeholder-Driven Model for Journal Policies." *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23336.
- Swanson, Alexandra, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. "Snapshot Serengeti, High-frequency Annotated Camera Trap Images of 40 Mammalian Species in an African Savanna." Dryad Digital Repository. doi:10.5061/dryad.5pt92.
- Tenopir, Carol, Ben Birch, and Suzie Allard. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*. An ACRL White Paper. Association of College and Research Libraries, a division of the American Library Association, 2012. [http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir\\_Birch\\_Allard.pdf](http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf).
- The Wellcome Trust. "Policy on Data Management and Sharing." Accessed August 6, 2016. <https://wellcome.ac.uk/funding/managing-grant/policy-data-management-and-sharing>.

- Thomson, Sara Day. "Technology Watch Report 16: Preserving Transactional Data." Digital Preservation Coalition. May 2, 2016. doi:10.7207/twr16-02.
- Tibbo, Helen R., and Christopher A. Lee. "Closing the Digital Curation Gap: A Grounded Framework for Providing Guidance and Education in Digital Curation." In *Archiving Conference*, vol. 2012, no. 1, pp. 57-62. Society for Imaging Science and Technology, 2012. <http://www.ils.unc.edu/caltee/p57-tibbo.pdf>.
- United States Government. "US Open Data Action Plan." May 9, 2014. [https://www.whitehouse.gov/sites/default/files/microsites/ostp/us\\_open\\_data\\_action\\_plan.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/us_open_data_action_plan.pdf).
- University of Illinois Urbana-Champaign School of Information Science. "Specialization in Data Curation." Accessed August 4, 2016. [http://www.lis.illinois.edu/academics/programs/specializations/data\\_curation](http://www.lis.illinois.edu/academics/programs/specializations/data_curation).
- University of Notre Dame. "About the eMotion and eCognition Lab." Accessed August 6, 2016. <http://www3.nd.edu/~emotecog/about.html>.
- US Geological Survey. "NBII to Be Taken Offline Permanently in January." *USGS Access Newsletter* 14, no. 3 (Fall 2011), [https://www2.usgs.gov/core\\_science\\_systems/Access/p1111-1.html](https://www2.usgs.gov/core_science_systems/Access/p1111-1.html).
- Van Noorden, Richard. "Irish University Labs Face External Audits." *Nature News*, June 17, 2014. <http://www.nature.com/news/irish-university-labs-face-external-audits-1.15422>.
- Vines, Timothy H., Arianne YK Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24, no. 1 (2014): 94-97. doi:10.1016/j.cub.2013.11.014.
- Witt, Michael. "Institutional Repositories and Research Data Curation in a Distributed Environment." *Library Trends* 57, no. 2 (2008): 191-201. doi:10.1353/lib.0.0029.
- Yahoo. "R10—Yahoo News Feed dataset, version 1.0 (1.5TB)." Accessed August 6, 2016. <http://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75>.





**PART I**  
**Setting the Stage for  
Data Curation**

*Policies, Culture, and  
Collaboration*









## CHAPTER 1\*

# Research and the Changing Nature of Data Repositories

*Karen S. Baker and Ruth E. Duerr*

## Introduction

This chapter explores the changing nature of research and data repositories. Trends in open data, big data, and long-tail data are ongoing,<sup>1</sup> following shifts from analog devices and documentation to digital instrumentation and digital data. Further, recent mandates about increasing access to data in the United States come at a time when digital capabilities are increasing though digital infrastructure is in flux.<sup>2</sup> Attention to and funding for data sharing have propelled data repository activities in both new and established digital settings. As the number and kind of repositories accepting research-generated data increase, their effectiveness depends upon developing widespread understanding of data concepts as well as the knowledge accumulated about successes and failures in the digital realm.

The full reality of managing research data and data repositories in a Digital Age is informed and shaped by past efforts carried out in many sectors. It is impacted by new participants, new roles, and changes in the distribution of responsibilities associated with data management. In addition, evolving technologies result in changing support mechanisms for documentation, preservation, and access of data. Contemporary data management efforts have more than fifty

---

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

years' experience to draw upon given early large-scale assemblies of digital data in scientific research fields such as remote sensing and weather as well as social science research fields such as survey and census methods.<sup>3</sup> Only a portion of the insights gained from past experience with data management and data systems are readily available given the combination of emphasis on scientific findings and of succinctness required in writing for the scholarly literature. Incentives and rewards for writing about work with data have been lacking.<sup>4</sup> New forums and journals are emerging that provide venues for discussions about past and present work with data so that past experience is available to new communities of data workers (see section "Changing Research Needs and New Initiatives" below).

This paper considers both conceptual and historical underpinnings in the story of data repositories. From work with data repositories in a variety of research fields, three concepts—data ecosystem, liaison work, and continuing design—help in understanding how work with digital data can contribute to the viability and well-being of the research process. These concepts, together with related issues and recommendations, are presented below as projects, communities, consortia, alliances, centers, programs, agencies, universities, publishers, libraries, and organizations of all kinds grapple with managing and preserving data in repositories.

## Background

A few early data efforts in the sciences are presented as examples of past activities that inform today's work.

### *Changing Support for Data*

Work with data is embedded in the processes, methods, and goals of research. Rigor in documenting thought processes, evidence collection, and data is integral to ensuring a robust research process. There is a long history of research data recorded in station books and laboratory notebooks.<sup>5</sup> In addition, white papers and project newsletters as well as expedition and technical reports full of tables of numbers were, and continue to be, published outside formal academic and commercial channels by a variety of organizations. Such materials, known as "the gray literature," are authoritative as primary sources. As the name suggests, however, they may be limited in terms of discoverability, access, and vetting. Nevertheless, these outlets have played a significant role in providing researchers access to data. While research findings traditionally appear in formal publication venues, the original, full data record was often in the gray literature as well as file cabinets.<sup>6</sup>

With the development of technologies such as cameras and strip chart recorders, a variety of organizational subunits such as photo labs emerged to

handle these analog materials and to support conversion to forms that could be published. Although they did not consider themselves data publishers, they or their counterparts routinely created reports with primary data in the form of tables, photos, maps, and graphs. Many of these offices have since closed or have been transformed, such as the photo lab that becomes a digital service group. Closing often occurred before infrastructure was in place to handle documentation and data in new ways beyond the capability of an individual's desktop. Eventually, with Internet availability, researchers and research groups developed new practices such as delivery of content including field data under a Data tab on a research website. In a sense, the current attention to data access and new forms of data citation is a return to the norm of retrieving and citing data that appeared in the print-based gray literature. With orders of magnitude more digital data generated, however, new kinds of digital tools, capabilities, and arrangements are required to support widespread access to digital data.

## *Expanding Support for Data in Natural and Social Sciences*

With the development of large-scale international research initiatives, support for data took a variety of forms. Spurred by twentieth-century post–World War II planning, a number of data facilities were established. For instance, World Data Centers and the Federation of Astronomical and Geophysical Data Analysis Services evolved, starting with the International Geophysical Year (IGY) in 1957–1958 with its focus on international science. From the IGY, a revolutionary vision of the earth as a whole emerged, focusing the attention of geoscientists collectively on scientific methods, measurements, and data. The International Council of Scientific Unions (now International Council for Science) established a system of World Data Centers to serve the IGY and developed data management plans for each IGY scientific discipline.<sup>7</sup> The World Data Centers focused on replicating data across the centers and sharing data across the globe. The ICSU Committee on Data for Science and Technology (CODATA) continues to develop and share knowledge about data today.<sup>8</sup> With their beginnings as centers full of the books and reports containing data for IGY and other initiatives, early data efforts grew to include magnetic tapes and punch cards at designated locations. Today management in data centers has grown to include digital data and physical samples as well as to accommodate many stakeholders and audiences.<sup>9</sup> The transition and renaming of the World Data Center system in 2009 to be the World Data System represents another shift in perspective with data envisioned within an interoperable set of systems.

In the United States, federal centers developed and took many forms. Federally Funded Research Development Centers (FFRDC) were created as public-private partnerships to support research community projects by making available large-scale resources such as the aircraft required for atmospheric science fieldwork.<sup>10</sup> Research support includes project coordination, instrumentation, field support, and work with data. National Data Centers such as the National Climate Data Center and the National Oceanographic Data Center were created in order to support management of data from platforms with large data streams such as from satellites. Supercomputer centers were developed as national resources to provide computational power to researchers across the nation.<sup>11</sup> These centers have developed repositories for data of many kinds existing alongside other preservation institutions such as archives with collections of photos and manuscripts, museums with physical artifacts, and libraries with books and journals. Tape racks proliferated as recordings on seven- and nine-track tapes replaced everything from strip chart recorders to images. Tapes were replaced in turn by new storage technologies. Many other, less visible changes were occurring in data centers, driven by changes in applications, configurations, budgets, institutions, and careers.<sup>12</sup> As the number of data centers grew, coordination activities started taking place. For instance, the National Archives and Records Administration (NARA) joined in 1992 with the scientific community and with federal and nonfederal entities that collect data about the earth to consider collectively data management and archiving procedures.<sup>13</sup> The ramifications of this interaction resulted in recommendations that NARA collaborate with other agencies that maintain long-term custody of data.

In the social sciences, early national-level repository development was spurred by an initial need for community access to data from election studies and from the US Census.<sup>14</sup> The Inter-university Consortium for Political and Social Research (ICPSR), which dates its origin to 1962, provides an example of responding to change over time. ICPSR began with a membership model to fund its data management costs but is now leading a call for change in support mechanisms for domain repositories.<sup>15</sup> This consortium has responded to community interests by participating in an alliance to distribute widely backup copies of data across several repositories. ICPSR has also responded to recent mandates for public data access by creating a new level of service. This service, called OpenICPSR, supports public availability of data free of cost.<sup>16</sup>

## *Data Repository Diversity*

Setting aside the issue of data presentation, we consider two categories of data repositories depending upon whether they ingest homogeneous or heterogeneous

types of data. *Data types* is an overloaded term; in this case, we are referring to sampling differences or their equivalent such as measurement and format differences. For example in the earth sciences, data is sampled in a variety of manners such as individual points of data, streamed data from a single location, and swaths or grids of data covering geographic areas. An early example of meeting large-scale, homogeneous data needs is satellite data managed by National Aeronautics and Space Administration (NASA) data facilities.<sup>17</sup> Similarly, the Protein Data Bank, a worldwide entity with portals that serve macromolecular structural data, handles highly structured data of a different sort.<sup>18</sup>

Data facilities that specialize in accepting homogeneous data are able to design a system crafted for organizing, preserving, and disseminating a particular type of data. By tailoring to a single type of data, a repository can provide more robust and advanced services for that data. Examples of advanced services include development of higher level data products, subsetting, rejections, aggregation services, and on-the-fly analysis. This is in contrast to data facilities that accept a broad range of data but are limited to providing that data back in a form similar to what was ingested. Recently, repositories such as Dryad and Figshare work with a wide variety of often less structured data objects associated with research rather than highly structured homogeneous data types.<sup>19</sup>

Assembling and organizing data highlights the differences in data and the need for a variety of data systems to support research.<sup>20</sup> A great deal is still to be learned from the diversity of data repositories—each developed with its particular goals, designers, developers, audience, time frame, infrastructure, workflow and products—whether dealing with homogeneous or heterogeneous data. Comparison of repository efforts provides insight into data management and system design. Dialogue across repositories is nascent, undergoing continuing development and building on experience from earlier data efforts. Registries of data repositories are adding to their visibility. Currently there exist more than a thousand repositories in the re3data registry of research repositories.<sup>21</sup> The kinds of data repositories are explored in the second volume of this two-volume set.<sup>22</sup>

## Three Concepts at Work

We present three concepts that relate to ongoing efforts to make research data available via repositories: data ecosystem, liaison work, and continuing design. These concepts support the changing work associated with research data. Whether embraced enthusiastically as challenges or acknowledged reluctantly as obstacles, these concepts address data issues occurring across a variety of settings. Additionally we suggest that they are central to ensuring that the development of data practices, processes, and systems is effective in supporting research.

## *Data Ecosystem: Growing Interdependence*

The concept of a data ecosystem is key because it fosters thinking about the interrelatedness of a multiplicity of repositories as well as activities associated with data in both the research and repository realms. With socio-technical insight, Parsons and colleagues defined a data ecosystem as “the people and technologies collecting, handling, and using the data and the interactions between them.”<sup>23</sup> Historically, much of the planning for data and data repositories occurred independently hidden behind laboratory, disciplinary, and commercial doors. Today there are increasing calls for open data,<sup>24</sup> and a growing tradition of describing data management and repository efforts in peer-reviewed journals.<sup>25</sup> With system architectures ranging from small-scale, custom designs to larger-scale systems with more generic, higher-level approaches, there are many choices to be made when data is assembled.

The concept of a data ecosystem captures the dynamics and feedbacks associated with data and data repositories. Work with data is impacted by short-term cycles such as project funding, field studies, experimental set-ups, and technology development. Work with data also involves longer-term influences such as research trends, institutional arrangements, career trajectories, and the growth of information infrastructure. Change within the ecosystem can occur due to any number of sudden events, including environmental disturbances, political shifts, or perhaps a human insight.

The interrelatedness of data itself blurs repository boundaries. For example, a repository may preserve a study that includes physical measurements, traditional knowledge, and artistic sensibilities. Indeed, what counts as data is in the eye of the beholder.<sup>26</sup> Data products from a single repository may be relevant to fields as diverse as natural sciences, humanities, and the arts, as illustrated by an example such as the data set NSIDC-0650.<sup>27</sup> Moreover, aggregating the data that results from small-scale, individual research projects often provides data products that are highly valued and used by broad communities. The Worldwide Protein Data Bank, the Interdisciplinary Earth Data Alliance, and the Linguistic Data Consortium are examples of this kind of repository.<sup>28</sup>

Within the data ecosystem, data is described using an array of metadata standards and minimum information guidelines.<sup>29</sup> The metadata contributes significantly to subsequent data discoverability, interpretation, and usability. Some standards are designed to facilitate computational functionality while others are streamlined to enable assembly of data at larger scales and broader scope. Decisions made about data documentation during the process of generation and curation determine what will be known about the data subsequently including its collection context relating to field circumstances and its research context relating

to why it was collected. Metadata choices involve planning for levels of descriptive completeness and are influenced by the availability of metadata validation techniques and crosswalks as well as the time available for documentation work. In time, new local metadata elements may evolve to describe particular aspects of the data at hand that are not addressed by an existing standard. For this reason, the use of not only standards but also working or local standards as well as participation by a wide range of data specialists in the standards-making process is critical.<sup>30</sup>

The activities within a repository, a subsystem within the data ecosystem, add to the dynamics of the system. They are represented as a data management stack of services defined by four layers: storage, archive, preservation, and curation. Figure 1.1 portrays the stack with the most basic level of service at the bottom.<sup>31</sup>

Layer #	Layer	Characteristics
4	Curation	Adding value throughout life cycle
3	Preservation	Ensuring that data can be fully used and interpreted
2	Archiving	Data protection including fixity and identifiers
1	Storage	Bits on disk, tape, cloud, etc.; back up and restore

**FIGURE 1.1**

Data management service stack model redrawn.<sup>32</sup>

---

Awareness of the full stack is needed if services meeting lower levels of functionality (layer 1) are to plan forward to achieve higher levels. Additional levels may be achieved either by internal expansion or by partnering with external service providers. One example of partnering is to expand by contracting for cloud-based storage as an addition to existing data system services. Another example would be a repository that creates a customized interface as the front end for a more standardized back end (e.g., Fedora or DSpace).\*

## *Liaison Work and Mediation*

In addition to the concept of a data ecosystem, liaison work is a second concept that adds to our understanding of change relating to work with research data. Liaison work involves consultation, mediation, advocacy, integration, synthesis, translation, and mutual learning. Support activities are carried out not only during deposit of data in a repository but throughout the research life cycle where assistance may be needed to initiate or support data practices and facilitate

---

\* For example, repositories using Islandora have a Drupal front-end “solution pack” with a Solr/Fedora back-end.

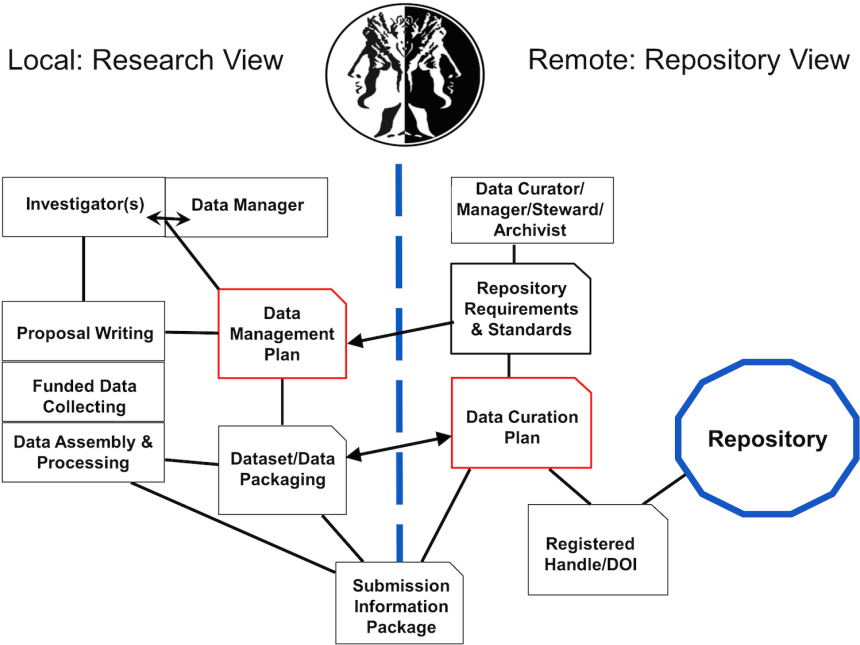


communication. Liaison work by data intermediaries may involve repackaging of data to facilitate new uses,<sup>33</sup> mediation to address some aspect of a shared information environment, migration of data to accommodate changes in technology or setting as well as the capture of data and metadata. Roles such as data manager, information manager, and data scientist are emerging in the local research realms of field and laboratory alongside the more traditional roles of technician, research assistant, and analyst. From a repository view of stakeholders, these positions are often considered as part of a broader category such as “scientist” or “researcher.”<sup>34</sup> Library subject specialists traditionally are referred to as liaisons, in this case focusing on interactions between the library and individuals who are experts in a field. Within libraries and repository arenas, however, new roles are also appearing with titles such as data specialist, data curator, and data steward. Organization charts, surveys, and the literature show the emergence of functional specialists such as digital services librarian, metadata librarian, scholarly communication specialist, and institutional repository coordinator, who work in collaboration with those who have established library titles associated with research data services.<sup>35</sup>

Responding to the need for a new data workforce, academic curricula are expanding to include certifications and degrees in data curation, informatics, information management, and data analytics at master’s and PhD levels.<sup>36</sup> Discussions, definitions, and clarifications of meaning are ongoing worldwide within and across institutions. Collective work is underway to catalog the various kinds of data concepts, data terms, and data repositories, thereby establishing common ground for those working with data.<sup>37</sup>

Perspectives on working with data vary depending upon one’s location within a data ecosystem. For example, two distinct arenas of work are displayed in figure 1.2: a local research view on the left and a repository view on the right. Differing perspectives are evident in the language used to refer to data sent to a repository data system: *data submission* is a phrase used by those outside the repository; in contrast, *data ingestion* refers to a formal system of accounting for and registering data within a repository. Attention to the research-repository interface requires time dedicated to communication and coordination. The distinction between the two views has been described as one involving differences in “sociotechnical distance” from the site of the data origin.<sup>38</sup> Workplace practices differ for some data-related tasks that may be carried out pre-submission or post-ingestion depending upon research-repository arrangements.

On the left side of figure 1.2, local data efforts include writing a data management plan (DMP) that outlines work with data prior to repository submission. An investigator who may work with colleagues from any number of departments, laboratories, projects, or libraries typically writes the DMP. This plan is part of the grant proposal writing process that, if successful, includes some funds for data collection as well as local data assembly, documenting, processing, and packag-



**FIGURE 1.2**

Two perspectives on working with data: on the left is a research view and on the right is a repository view.

ing. When the writers of the DMP identify a repository to which they plan to submit their data, repository requirements may inform their data management plan as is indicated in figure 1.2 by the unidirectional arrow. Not all data is processed or analyzed for final use. The majority of accessible research data made available over the last decades in repositories has been identified and selected by a research community as the most significant data for their particular interests as resource and reference data sets.<sup>39</sup> Data packaging for submission is given a box in figure 1.2 in order to highlight the decision making and other work entailed in identifying data for submission as well as the work involved in creating final data sets that are formatted and documented according to repository requirements.

On the right, repository efforts center on a data curation plan (DCP) that describes the planned procedures within a data repository. These are the tasks required to ingest, archive, and make accessible the data. There are typically steps for identifying, collecting, formatting, programming, and documenting the data in preparation for access. Also illustrated in figure 1.2 are repository steps taken in assigning a unique identifier, either a local handle or a global identifier such as a digital object identifier (DOI).<sup>40</sup> Repositories differ in post-ingest procedures;

some perform no curation, while others require conformance with repository requirements before data sets are released into the repository catalog or made available at the user interface. Sometimes there is reciprocal communication between researchers and repository staff such that each informs and influences the other as collaborators in activities that may include identifying data-specific troubles, addressing community issues, and developing new vocabularies.

## *Continuing Design: Standards, Systems, and Models*

In work with digital data, change is captured not only in terms of a complex data ecosystem and by liaison work, but also by a third concept of “continuing design.” Within a continuing design environment, data activities involving terminology and procedures, as well as workflows and systems, change over time. Continuing design describes an adaptive strategy of expecting change; it is an approach to work where the goal of planning for a “final solution” is reconceived as continually taking into account new or related factors that inform iterative cycles of redesign. From a perspective that nothing is permanent and that change is inevitable, making plans to adapt becomes second nature. Continuing design is reported in practice as carried out in both incremental steps and breakthrough improvements.<sup>41</sup> This approach to design follows earlier work on “continuing design in use” in the field of information systems.<sup>42</sup> There is recent interest in action-centric approaches to design. For instance, agile development is a technique adopted by some system developers, and more recently, agile curation is being considered in addressing approaches to data activities.<sup>43</sup>

Examples of continuing design exist for cases of metadata, systems, and models. Development of metadata incorporating community vocabularies or shared ontologies illustrate iterative development over time. In the social sciences, the international community coalesced early around the Data Documentation Initiative (DDI) standard.<sup>44</sup> In the earth sciences, the NASA Directory Interchange Format (DIF) evolved to encompass all US agencies to become the Content Standard for Digital Geospatial Metadata (CSDGM) through the auspices of the Federal Geographic Data Committee (FGDC). It finally gained international standing as the ISO 19115 family of metadata standards. In the case of biodiversity data, Darwin Core began development based on the standards established by the Dublin Core Metadata Initiative and continues to evolve using a working group model.<sup>45</sup>

The Earth Observing System Data and Information System (EOSDIS) provides an example of an early, foundational system that has changed over time.<sup>46</sup> Developed in 1986 for NASA’s Earth Science Data Systems, it brought together distributed data holdings from multiple repositories in a single interface. Cur-

rently this actively evolving system supports over 2 million users a year and delivers on average 27.9 TB of data each day.<sup>47</sup> A final example of continuing design for satellite data is in managing and archiving large streams of data across many countries. The Consultative Committee for Space Data Systems (CCSDS) was formed in 1982 as a multinational organization of space agencies to coordinate data work and recommend standards. Though originating in a multinational satellite community, the Reference Model for an Open Archival Information System (OAIS) has had extensive uptake by many other communities. The OAIS conceptual framework represents an early, effective case of large-scale coordination work based on over fifty years' experience in trying to make data usable over time.<sup>48</sup> The level of representation is such that it facilitates communication across various communities and contexts yet still allows for differences in data systems. The OAIS definitions of terms such as *information system*, *designated community*, and *information package* have proven extremely useful in many data arenas.

## Changing Research Needs and New Initiatives

In an effort to ensure that research is meeting the needs of society, the Office of Science and Technology mandated that the results of federally funded research, including data, be publicly accessible.<sup>49</sup> This led to agencies changing their data policies and requirements. One might consider a mandate to share research data before tools and infrastructure are in place as an example of the cart before the horse. The responses evident today suggest otherwise. For instance, the requirement by a wide range of funding agencies for a data management plan (DMP) to accompany research proposals serves as an effective first step in creating data management awareness and dialogue from individual researchers to organizational and agency management.<sup>50</sup> When DMPs are aggregated and mined, they provide overviews of what arrangements are being made in practice in a wide variety of circumstances, thereby providing feedback about existing and imagined services.<sup>51</sup> These plans document individual understandings and local actions as well as revealing misconceptions about data access and data systems.

Researchers and repository staff alike require time to pilot and gain experience with the new realities of the call for open data and its impact on research, reference, and resource collections of data. Researchers must update old beliefs such as “I can’t stop to document my data but someone else can do it later.”<sup>52</sup> And repositories must update views such as “We can create services, and they will be used.”<sup>53</sup> During this interim period, as data becomes available and reuse increases, researchers may see the value of other people sharing their data but may still believe that they are not required to do the same.<sup>54</sup> The three concepts of data

ecosystem, liaison work, and continuing design facilitate a fuller understanding of work with research data during this time of transition. In time, digital capabilities and services will mature and become part of the information infrastructure.

Liaison work is related to some of the social aspects of information infrastructure, including awareness of issues related to the responsible conduct of research.<sup>55</sup> The 2009 NAP report discussed the role of liaisons in terms of data professionals working with researchers because “As new methods and tools are brought into practice, researchers are continually challenged to understand them and use them effectively.”<sup>56</sup> Whose responsibility is writing about the practical knowledge, knowledge that goes beyond what is typically found in metadata documentation? Is it the research scientist, a data scientist, or an emergent liaison position? This question becomes more complicated when the role of technology and technologists is included. When responsibility for work with data is delegated to data professionals, then support is needed for their work as well as acknowledgment for their intellectual contributions. Traditionally, data professionals are not funded to document their work separately from the research the documentation supports. Further discussion and new arrangements will be needed to ensure support for not only research data but also the full documentation of methods and limitations of publicly reported data.

The library community provides a valuable, publicly accessible model of interrelated services that expands beyond geographic territories and hierarchies to support both overlapping activities and special niche services. Library infrastructure illustrates arrangements that foster outreach to library users as well as in-reach among library professionals. Unlike the case for many data professionals in research arenas, library professionals have established a mature form of infrastructure complete with a variety of forms of communication ranging from working groups and conferences to surveys, reports, and journals. With many institutional repositories being developed and maintained in conjunction with libraries,<sup>57</sup> now is a good time to consider this question: How will library data professionals contribute to and coordinate with the broader ecosystem of data repositories?

In terms of the technical aspects of information infrastructure, the concept of continuing design imparts the need to anticipate change for each standard, data system, and model. As an example, brokers are one of the recent responses to the need for interactions across various work arenas. Much like a currency exchange station where you can change your money to that of another country, brokers translate across differences.<sup>58</sup> They negotiate across heterogeneous data and metadata formats as well as types of services. For instance, using machine-to-machine communication, they translate a local metadata format to a variety of other formats used in other settings. Brokering systems are middleware that provide real-time mediation between machines.<sup>59</sup> They use programming techniques to patch across communities that have different standards and formats. Translation

may involve creation of a more general product such a derived data set, for example, an average or synthesis from the original digital formulation.

Open access within a large-scale data ecosystem, while an ideal goal, is a nuanced concept. The phrase “ethically open access” is used in cases where not all research data can be made open. The reasons for a qualified openness include the need for laws and regulations that protect privacy, security, and legitimate commercial and community interests (e.g., endangered species, archeological sites). In addition, there are the ethics of dealing with local and traditional knowledge where the data is not the property of the researcher but is instead the property of the knowledge holder.<sup>60</sup> These issues require policy development and impact the continuing design of data systems.

New forums are part of the infrastructure emerging to address coordination and communication within the data ecosystem. The collective knowledge building carried out in these venues facilitates integrative work with data across institutions and sectors. Some repositories focus exclusively on institutionally specific materials. Many institutional repositories, however, have only recently begun to work with and characterize their data holdings. Some aim to include the publications of members in the organization. In the United States, one response, along traditional lines, to the call for opening access to the results of publicly funded research, is the Clearinghouse for the Open Research of the United States (CHORUS). CHORUS, established as a nonprofit cooperative effort to coordinate services in scholarly publishing for public benefit, involves publishers, funding agencies, and technology and resource partners.<sup>61</sup> A library-community response to open access is the Shared Access Research Ecosystem (SHARE), with participants including the Association of American Universities, the Association of Public and Land-Grant Universities, and the Association of Research Libraries.<sup>62</sup> The Confederation of Open Access Repositories (COAR) is a recently formed international effort of institutional repositories focused on theses and papers.<sup>63</sup> Another recent initiative is the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS), which brings together domain data facilities and publishers with a strategy for developing relations between publishers of journals and select repositories certified to curate data associated with publications.<sup>64</sup>

The forums serve as neutral venues for development of a wide variety of integrative practices, including development of data citation practices and of the NASA Earth Science Data Preservation Content Specification.<sup>65</sup> The Earth Science Information Partnership (ESIP) is a disciplinary forum that reaches across agencies and communities.<sup>66</sup> With data repositories existing worldwide, recognition and support of the data ecosystem as a global affair is provided by new kinds of partnering, such as the Research Data Alliance (RDA).<sup>67</sup> RDA, an international and interdisciplinary forum for data and research professionals, supports a number of active interest groups and working groups. Its strategic goals aim to bridge international and disciplinary boundaries by enhancing interopera-

bility. Such forums facilitate communication among stakeholders in new ways. They prioritize inclusivity, recognizing that a diversity of perspectives is crucial to broadening and deepening our understanding of data and repository work.

## Final Thoughts

The mandate to share and provide ethically open access to data is galvanizing change in data practices in the realms of research and repositories. The drive to provide data access and to enhance research capabilities is leading to the development of new concepts, roles, and ways of working. The data ecosystem conveys the complexity and scale of the setting where integrative work in social, technical, organizational, and political realms is needed. The flexibility of the data ecosystem accommodates both heterogeneity and standardization. Data-sharing practices are unfolding, informed by a diversity of research efforts that engage a variety of participants. New forums for communication come at a time when data discussions will benefit from a broad set of voices. A loosely coupled data ecosystem, facilitated by liaison and continuing design work, provides an environment amenable to change as well as diversity.

We must continue to ask probing questions: When data is both more readily available and provided in forms amenable to multiple audiences, are a variety of environmental, social, and economic issues in research and public arenas addressed more effectively? Will change result in loss of data or of scientific innovation? Can risks associated with sustainability of access be minimized? Access is dependent upon informed decision making about levels of selection and preservation. Awareness of the dynamics and multiplicity of elements in the data ecosystem is critical, while maintaining the capacity to provide services for a diversity of data arrangements is a challenge. Sustainability and risk-of-loss issues are evident regardless of where repositories reside institutionally: for example, in national centers, academic environments, multi-site consortia, or ad hoc collections in a laboratory.<sup>68</sup> Some combinations of sustainability and risk may lead to—some might say lead further into—a digital dark age.<sup>69</sup>

The authors, who draw on activities described above and on personal experience with data projects primarily in the sciences, formulate a few basic recommendations for participants in the ecosystem of data and repositories:

- Identify and incorporate lessons from the past.
- Recognize the importance of a variety of data repositories.
- Plan for essential services and definitions of basic infrastructure to expand.
- Consider how to capture the documentation needed for sharing data.
- Recognize the dynamics of local-scale and large-scale data efforts in the data ecosystem.



- Value loosely structured as well as highly structured information environments.
- Recognize the role of data professionals and other liaisons in research support.

Repositories are, and need to be, works in progress in order to be responsive to the ongoing change that is integral to research and work with data. Not only individual researchers but those working with data and the development of information infrastructure require latitude “to follow hunches, experiment with methods, explore conjectures, and make mistakes.”<sup>70</sup> In some situations, an overemphasis on repository ingestion and holdings or on compliance enforcement may prove counterproductive. Though standardization facilitates ease of data reuse and assessment, due diligence is required in contemporary data environments to avoid a “one size fits all” administration that unduly constrains research and the dynamics of responding to change. Research, in a continuing quest for knowledge, is about discovering the unknown. Data is central to the research effort. Today, changing expectations and capabilities with digital data impact the entire research process at a time our understanding of information infrastructure and data repositories is nascent. In the ongoing transition to the Digital Age, research requires active partnership with a variety of kinds of data repositories.

## Notes

1. Virginia Gewin, “Data Sharing: An Open Mind on Open Data,” *Nature* 529, no. 7584 (2016): 117–19; Royal Society Science Policy Centre, *Science as an Open Enterprise*, final report (London: The Royal Society, June 2012), <https://royalsociety.org/-/media/policy/projects/sape/2012-06-20-saoe.pdf>; Peter Fox and Ray Harris, “ICSU and the Challenges of Data and Information Management for International Science,” *Data Science Journal* 12 (2013): WDS1–WDS12, doi:10.2481/dsj.WDS-001; Tony Hey, Stewart Tansley, and Kristin Tolle, eds., *The Fourth Paradigm*, vol. 1 (Redmond, WA: Microsoft Research, 2009), <http://research.microsoft.com/en-us/collaboration/fourth-paradigm/contents.aspx>; P. Bryan Heidorn, “Shedding Light on the Dark Data in the Long Tail of Science,” *Library Trends* 57, no. 2 (2008): 280–99.
2. John P. Holdren, “Increasing Access to the Results of Federally Funded Scientific Research,” Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013, [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
3. Ruth Duer, “Data Archives and Repositories,” in *Encyclopedia of Remote Sensing*, ed. Eni G. Nijoku (New York: Springer, 2014), 127–31; Myron Gutmann, Kevin Schürer, Darrell Donakowski, and Hilary Beedham, “The Selection, Appraisal, and Retention of Social Science Data,” *Data Science Journal* 3 (2006): 209–21, doi:10.2481/dsj.3.209.
4. National Academy of Sciences, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age* (Washington DC: National Academies Press, 2009), doi:10.17226/12615.



5. Robert R. Downs, Ruth Duerr, Denise Hills, and Hampapuram K. Ramapriyan, "Data Stewardship in the Earth Sciences," *D-Lib Magazine* 21, no. 7/8 (July/August 2015): 2, doi:10.1045/july2015-downs.
6. Ibid.
7. International Council for Science, *Ad Hoc Strategic Committee on Information and Data: Final Report to the ICSU Committee on Scientific Planning and Review* (Paris: International Council for Science, 2008), <http://www.icsu.org/publications/reports-and-reviews/scid-report>; International Council of Scientific Unions Panel on World Data Centres, *Guide to the World Data Center System*, (Paris: International Council of Scientific Unions, 1996), <http://www.ukssdc.ac.uk/wdc/guide/wdcguide.pdf>.
8. CODATA (Committee on Data for Science and Technology) homepage, accessed May 17, 2016, <http://www.codata.org>.
9. E.g., Fox and Harris, "ICSU and the Challenges of Data"; Karen S. Baker, Ruth E. Duerr, and Mark A. Parsons, "Knowledge Mobilization: Co-evolution of Data Products and Designated Communities," *International Journal of Digital Curation*, in press.
10. Jill M. Hruby, Dawn K. Manley, Ronald E. Stoltz, Erik K. Webb, and Joan B. Woodard, "The Evolution of Federally Funded Research and Development Centers," *Public Interest Report* 64, no. 1 (Spring 2011): 24–31, <http://www.fas.org/pubs/pir/2011spring/FFRDCs.pdf>.
11. Susan L. Graham, Marc Snir, and Cynthia A. Patterson, *Getting up to Speed: The Future of Supercomputing* (Washington, DC: National Academies Press, 2004), <http://www.nap.edu/catalog/11148/getting-up-to-speed-the-future-of-supercomputing>.
12. National Research Council, *Government Data Centers* (Washington, DC: National Academies Press, 2003), doi:10.17226/10664.
13. National Research Council, *Preserving Scientific Data on Our Physical Universe* (Washington, DC: National Academies Press, 1995), doi:10.17226/4871.
14. Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, eds., *Sharing Research Data* (Washington, DC: National Academies Press, 1985), doi:10.17226/2033.
15. Inter-university Consortium for Political and Social Research, "Sustaining Domain Repositories for Digital Data: A Call for Change from an Interdisciplinary Working Group of Domain Repositories," June 24–25, 2013, <http://www.icpsr.umich.edu/icpsr-web/ICPSR/support/announcements/2013/09/sustaining-domain-repositories-for>.
16. OpenICPSR homepage, accessed May 18, 2016, <https://www.openicpsr.org>.
17. National Research Council, *Review of NASA's Distributed Active Archive Centers*, (Washington, DC: National Academies Press, 1999), doi:10.17226/6396.
18. Helen M. Berman, "The Protein Data Bank: A Historical Perspective," *Acta Crystallographica Section A: Foundations of Crystallography* 64, no. 1 (2008): 88–95.
19. Jane Greenberg, Hollie C. White, Sarah Carrier, and Ryan Scherle, "A Metadata Best Practice for a Scientific Data Repository," *Journal of Library Metadata* 9, no. 3–4 (2009): 194–212; Figshare homepage, accessed May 18, 2016, <https://figshare.com>.
20. Matthew S. Mayernik, "Research Data and Metadata Curation as Institutional Issues," *Journal of the Association for Information Science and Technology* 67, no. 4 (April 2016): 973–93, doi:10.1002/asi.23425; Research Information Network and the British Library, *Patterns of Information Use and Exchange* (London: Research Information Network, 2009), [http://www.rin.ac.uk/system/files/attachments/Patterns\\_information\\_use-REPORT\\_Nov09.pdf](http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf); Michael Witt, Jacob Carlson, D. Scott Brandt, and Melissa H. Cragin, "Constructing Data Curation Profiles," *International Journal of*

- Digital Curation* 4, no. 3 (2009): 93–103; for example, see NSF EarthCube-sponsored Roadmaps, accessed May 18, 2016, <http://earthcube.org/type-document/roadmaps>; Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, Recommended Practice CCSDS 650.0-M-2, Magenta Book, Issue 2 (Washington, DC: CCSDS Secretariat, June 2012), <http://public.ccsds.org/publications/archive/650x0m2.pdf>.
21. Re3data (registry of Research Data Repositories) homepage, accessed May 18, 2016, <http://re3data.org>; Heinz Pampel, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf, “Making Research Data Repositories Visible: The re3data.org Registry,” *PLOS ONE* 8, no. 11 (2013): e78080, doi:10.1371/journal.pone.0078080.
  22. Karen S. Baker and Ruth E. Duerr, “Data and the Diversity of Repositories,” in *Curating Research Data, Volume 2: A Handbook of Current Practice*, ed. Lisa R. Johnston (Chicago: Association of College and Research Libraries, 2016), 141–145.
  23. Mark A. Parsons, Oystein Godøy, Ellsworth LeDrew, Taco F. De Bruin, Bruno Danis, Scott Tomlinson, and David Carlson, “A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science,” *Journal of Information Science* 37, no. 6 (2011): 557.
  24. Rufus Pollock, “Building the (Open) Data Ecosystem,” *Open Knowledge Foundation Blog*, March 31, 2011, <http://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem>; Teresa M. Harrison, Theresa A. Pardo, and Meghan Cook, “Creating Open Government Ecosystems: A Research and Development Agenda,” *Future Internet* 4, no. 4 (2012): 900–928; Christine L. Borgman, “The Conundrum of Sharing Research Data,” *Journal of the American Society for Information Science and Technology* 63, no. 6 (2012): 1059–78.
  25. E.g., Mark A. Parsons and Ruth Duerr, “Designating User Communities for Scientific Data: Challenges and Solutions,” *Data Science Journal*, no. 4 (2006): 31–38, doi:10.2481/dsj.4.31; William K. Michener, John Porter, Mark Servilla, and Kristin Vanderbilt, “Long Term Ecological Research and Information Management,” *Ecological Informatics* 6, no. 1 (January 2011): 13–24; Michael Witt, “Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service,” *Journal of Library Administration* 52, no. 2 (2012): 172–88; Matthew S. Mayernik, Tim DiLauro, Ruth Duerr, Elliot Metsger, Anne E. Thessen, and G. Sayeed Choudhury, “Data Conservancy Provenance, Context, and Lineage Services: Key Components for Data Preservation and Curation,” *Data Science Journal* 12 (2013): 158–71, doi:10.2481/dsj.12-039; Stanislav Pejša, Shirley J. Dyke, and Thomas J. Hacker, “Building Infrastructure for Preservation and Publication of Earthquake Engineering Research Data,” *International Journal of Digital Curation* 9, no. 2 (2014): 83–97; Daryl E. Herzmann, Lori J. Abendroth, and Landon D. Bunderson, “Data Management Approach to Multidisciplinary Agricultural Research and Syntheses,” *Journal of Soil and Water Conservation* 69, no. 6 (2014): 180A–185A; Constanze Curdt and Dirk Hoffmeister, “Research Data Management Services for a Multidisciplinary, Collaborative Research Project,” *Program: Electronic Library and Information Systems* 49, no. 4 (2015): 494–512, doi:10.1108/PROG-02-2015-0016.
  26. Christine L. Borgman, *Big Data, Little Data, No Data* (Cambridge, MA: MIT Press, 2015).
  27. Shari Gearheard, *When the Weather Is Uggianaqtuq: Inuit Observations of Environmental Changes*, version 1, data set NSIDC 0650 (Boulder, CO: NSIDC: National Snow and

- Ice Data Center, 2004), accessed March 3, 2016, <https://nsidc.org/data/nsidc-0650>; Shari Gearheard, Matthew Pocernich, Ronald Stewart, Joelle Sanguya, and Henry P. Huntington, "Linking Inuit Knowledge and Meteorological Station Observations to Understand Changing Wind Patterns at Clyde River, Nunavut," *Climatic Change* 100, no. 2 (2010): 267–94.
28. Worldwide Protein Data Bank homepage, accessed May 18, 2016, <http://www.wwpdb.org>; Interdisciplinary Earth Data Alliance homepage, accessed May 18, 2016, <http://www.iedadata.org>; Linguistic Data Consortium homepage, accessed May 18, 2016, <https://www.ldc.upenn.edu>.
  29. "List of Metadata Standards," Joint Information Systems Committee, Digital Curation Center, accessed March 3, 2016, <http://www.dcc.ac.uk/resources/metadata-standards/list>; see Biosharing.org, accessed March 3, 2016, <http://biosharing.org>, for registry of standards, databases, policies, and collections in the life, environmental, and biomedical sciences; for visualization of metadata specifications, see Jenn Riley, "Seeing Standards: A Visualization of the Metadata Universe," poster, 2010, accessed March 3, 2016, <http://jennriley.com/metadatamap/>.
  30. Lynn Yarmey and Karen S. Baker, "Towards Standardization: A Participatory Framework for Scientific Standard-Making," *International Journal of Digital Curation* 8, no. 1 (2013): 157–72.
  31. "Sayeed Choudhury on Data Stack Model," YouTube video, 5:53, posted by CLIRDLF, August 7, 2012, <https://www.youtube.com/watch?v=3MD7KjZF34Y>; Mayernik et al., "Data Conservancy Provenance."
  32. Matthew S. Mayernik, Tim DiLauro, Ruth Duerr, Elliot Metsger, Anne E. Thessen, and G. Sayeed Choudhury, "Data Conservancy Provenance, Context, and Lineage Services: Key Components for Data Preservation and Curation," *Data Science Journal* 12 (2013): 161, doi:10.2481/dsj.12-039.
  33. Baker, Duerr, and Parsons, "Knowledge Mobilization."
  34. Liz Lyon, *Dealing with Data*, Consultancy Report, University of Bath, UK: UKOLN, 2007), [http://opus.bath.ac.uk/412/1/dealing\\_with\\_data\\_report%2Dfinal.pdf](http://opus.bath.ac.uk/412/1/dealing_with_data_report%2Dfinal.pdf).
  35. Janice M. Jaguszewski and Karen Williams, *New Roles for New Times* (Washington, DC: Association of Research Libraries, August 2013), <http://www.arl.org/component/content/article/6/2893>; Mark P. Newton, Christopher C. Miller, and Marianne Stowell Bracke, "Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force," *Collection Management* 36, no. 1 (2010): 53–67; Michael Witt, "Institutional Repositories and Research Data Curation in a Distributed Environment," *Library Trends* 57, no. 2 (2008): 191–201; Sarah Jones, Graham Pryor, and Angus Whyte, *How to Develop Research Data Management Services: A Guide for HEIs*, DCC How-to Guides (Edinburgh: Digital Curation Center, 2013), <http://hdl.handle.net/11329/250>.
  36. National Research Council, *Preparing the Workforce for Digital Curation* (Washington, DC: National Academies Press, 2015), doi:10.17226/18590.
  37. Data Foundation and Terminology Working Group, Research Data Alliance, accessed May 18, 2016, <https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html>.
  38. Karen S. Baker and Lynn Yarmey, "Data Stewardship: Environmental Data Curation and a Web-of-Repositories," *International Journal of Digital Curation* 4, no. 2 (2009): 12–27.
  39. National Science Board, *Long-Lived Digital Data Collections* (Arlington, VA: National Science Foundation, September 2005).

40. Joan Starr, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, et al., "Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications," *PeerJ Computer Science* 1 (May 27, 2015): e1, doi:10.7717/peerj-cs.1.
41. Helena Karasti, Karen S. Baker, and Florence Millerand, "Infrastructure Time: Long-Term Matters in Collaborative Development," *Computer Supported Cooperative Work (CSCW)* 19, no. 3–4 (2010): 377–415.
42. Austen Henderson and Morten Kyng, "There's No Place Like Home: Continuing Design in Use," in *Design at Work: Cooperative Design of Computer Systems*, ed. Joan Greenbaum and Morten Kyng, (Hillsdale, NJ: Lawrence Erlbaum Associates, 1991), 219–40.
43. Josh Young, W. Christopher Lenhardt, Mark Parsons, and Karl Benedict, "Taking Another Look at the Data Management Life Cycle: Deconstruction, Agile, and Community," poster, in *AGU Fall 2014 Meeting Abstracts*, Vol. 1 (Washington, DC: American Geophysical Union, 2014), 3779, accessed July 15, 2016, <http://abstractsearch.agu.org/meetings/2014/FM/IN51B-3779.html>.
44. Mary Vardigan, "The DDI Matures: 1997 to Present," *IASSIST Quarterly* 37, no. 1–4 (Spring 2013): 45.
45. Darwin Core Task Group, "Darwin Core," Biodiversity Information Standards (TDWG), last modified June 5, 2015, <http://rs.tdwg.org/dwcl>.
46. See Earthdata, EOSDIS (Earth Observing System Data and Information System), NASA, accessed May 18, 2016, <https://earthdata.nasa.gov/>.
47. H. K. Ramapriyan, "The Role and Evolution of NASA's Earth Science Data Systems" (slides for presentation at IEEE EDS/CAS Chapter, Camarillo, CA, August 19, 2015), last accessed February 13, 2016, <http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150018076.pdf>.
48. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*.
49. Holdren, "Increasing Access to the Results"; Timothy M. Beardsley, "Notes on Changing Practices in Data Publication," *BioScience* 65, no. 7 (2015): 645–50.
50. Amanda Whitmire, Kristin Briney, Amy Nurnberger, Margaret Henderson, Thea Atwood, Margaret Janz, Wendy Kozlowski, Sherry Lake, Micah Vandegrift, and Lisa Zilinski, "A Table Summarizing the Federal Public Access Policies Resulting from the US Office of Science and Technology Policy Memorandum of February 2013," v. 5, April 18, 2016, Figshare, doi:10.6084/m9.figshare.1372041.
51. Lizzy Rolando, Jake Carlson, Patricia Hswe, Susan Wells Parham, Brian Westra, and Amanda L. Whitmire, "Data Management Plans as a Research Tool," *Bulletin of the American Society for Information Science and Technology* 41, no. 5 (2015): 43–45; Carolyn Bischoff and Lisa R. Johnston, "Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University," *Journal of Librarianship and Scholarly Communication* 3, no. 2 (2015): eP1231, doi:10.7710/2162-3309.1231.
52. Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame, "Data Sharing by Scientists: Practices and Perceptions," *PLOS One* 6, no. 6 (2011): e21101, doi:10.1371/journal.pone.0021101.
53. G. Sayeed Choudhury, "Case Study in Data Curation at Johns Hopkins University," *Library Trends* 57, no. 2 (Fall 2008): 211–20; Dorothea Salo, "Innkeeper at the Roach Motel," *Library Trends* 57, no. 2 (Fall 2008): 98–123.

54. Carol Tenopir, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett, "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide," *PLOS ONE* 10, no. 8 (2015): e0134826, doi:10.1371/journal.pone.0134826.
55. David B. Resnik and Gregg E. Dinse, "Do US Research Institutions Meet or Exceed Federal Requirements for Instruction in Responsible Conduct of Research? A National Survey," *Academic Medicine* 87, no. 9 (2012): 1237.
56. National Academy of Sciences, *Ensuring the Integrity*, 60.
57. Clifford A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL Bimonthly Report* 226 (February 2003): 1–7, <http://www.cni.org/publications/cliffs-pubs/institutional-repositories-infrastructure-for-scholarship/>; Tyler O. Walters, "The New Academic Library: Building Institutional Repositories to Support Changing Scholarly and Research Processes," in *Proceedings of the ACRL Thirteenth National Conference*, ed. Hugh A. Thompson (Chicago: American Library Association, 2007), 56–63, <http://www.ala.org/acrl/files/conferences/confsandpreconfs/national/baltimore/papers/56.pdf>; Association of Research Libraries, *The Research Library's Role in Digital Repository Services*, Final Report of the ARL Digital Repository Issues Task Force (Washington, DC: Association of Research Libraries, 2009), <http://www.arl.org/storage/documents/publications/repository-services-report-jan09.pdf>; Sarah L. Shreeves and Melissa H. Cragin, "Introduction: Institutional Repositories: Current State and Future," *Library Trends* 57, no. 2 (2008): 89–97.
58. Stefano Nativi, Max Craglia, and Jay Pearlman, "Earth Science Infrastructures Interoperability: The Brokering Approach." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6, no. 3 (2013): 1118–29.
59. For example, see the US Geoscience Information Network Commons (GI-Cat), <http://lab.usgin.org/applications/gi-cat>; EarthCube Brokers, <http://earthcube.org/group/bcube>; Jeff McWhirter, Christopher J. Crosby, Peter C. Griffith, Siri Khalsa, Matthew A. Lazzara, and W. Jeff Weber, "Rich Support for Heterogeneous Polar Data in RAMADDA," in *AGU Fall Meeting Abstracts*, vol. 1 (Washington, DC: American Geophysical Union, 2013), 1632, <http://www.unidata.ucar.edu/committees/strat-com/2008spring/statusreports/ramadda.html>.
60. Francesco Mauro and Preston D. Hardison, "Traditional Knowledge of Indigenous and Local Communities: International Debate and Policy Initiatives," *Ecological Applications* 10, no. 5 (2000): 1263–69; Peter Pulsifer, Shari Gearheard, Henry P. Huntington, Mark A. Parsons, Christopher McNeave, and Heidi S. McCann, "The Role of Data Management in Engaging Communities in Arctic Research: Overview of the Exchange for Local Observations and Knowledge of the Arctic (ELOKA)," *Polar Geography* 35, no. 3–4 (2012): 271–90; Ryan K. Brook and Stéphanie M. McLachlan, "Trends and Prospects for Local Knowledge in Ecological and Conservation Research and Monitoring," *Biodiversity and Conservation* 17, no. 14 (2008): 3501–12.
61. CHORUS (Clearinghouse for the Open Research of the United States) homepage, accessed June 15, 2016, <http://www.chorusaccess.org>.
62. Association of Research Libraries, "SHARE (Shared Access Research Ecosystem)," accessed June 15, 2016, <http://www.arl.org/focus-areas/shared-access-research-ecosystem-share#.Vtj6Gdy8vt8>; Tyler Walters and Judy Ruttenberg, "Shared Access Research Ecosystem," *Educause Review* 49, no. 2 (2014): 56–57.
63. COAR (Confederation of Open Access Repositories) homepage, accessed June 15, 2016, <https://www.coar-repositories.org>.

64. COPDESS (Coalition on Publishing Data in the Earth and Space Sciences) homepage, accessed June 15, 2016, <http://www.copdess.org>.
65. American Geophysical Union, "AGU Publications Data Policy," accessed June 15, 2016, <http://publications.agu.org/author-resource-center/publication-policies/data-policy/>; NASA Earthdata, "NASA Earth Science Data Preservation Content Specification," accessed June 15, 2016, <https://earthdata.nasa.gov/standards/preservation-content-spec>.
66. Downs et al., "Data Stewardship in the Earth Sciences," 2.
67. Mark A. Parsons and Francine Berman, "The Research Data Alliance: Implementing the Technology, Practice and Connections of a Data Infrastructure," *Bulletin of the American Society for Information Science and Technology* 39, no. 6 (2013): 33–36.
68. Francine Berman and Vint Cerf, "Who Will Pay for Public Access to Research Data?" *Science* 341, no. 6146 (2013): 616–17.
69. Terry Kuny, "The Digital Dark Ages? Challenges in the Preservation of Electronic Information," *International Preservation News*, no. 17 (1998): 8–13.
70. National Academy of Sciences, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*, Washington DC: National Academies Press, 2009, doi:10.17226/12615.

## Bibliography

- American Geophysical Union. "AGU Publications Data Policy." Accessed June 15, 2016. <http://publications.agu.org/author-resource-center/publication-policies/data-policy/>.
- Association of Research Libraries. "SHARE (Shared Access Research Ecosystem)." Accessed June 15, 2016. <http://www.arl.org/focus-areas/shared-access-research-ecosystem-share#VtJ6Gdy8vt8>.
- . *The Research Library's Role in Digital Repository Services*, Final Report of the ARL Digital Repository Issues Task Force. (Washington, DC: Association of Research Libraries, 2009). <http://www.arl.org/storage/documents/publications/repository-services-report-jan09.pdf>.
- Baker, Karen S., and Ruth E. Duerr. "Data and a Diversity of Repositories." In *Curating Research Data, Volume Two: A Handbook of Current Practice*, edited by Lisa R. Johnston, 141–145 Chicago: Association of College and Research Libraries, 2016.
- Baker, Karen S., Ruth E. Duerr, and Mark A. Parsons. "Knowledge Mobilization: Co-evolution of Data Products and Designated Communities." *International Journal of Digital Curation* 10, no. 2 (2015). doi:10.2218/ijdc.v10i2.346.
- Baker, Karen S., and Lynn Yarmey. "Data Stewardship: Environmental Data Curation and a Web-of-Repositories." *International Journal of Digital Curation* 4, no. 2 (2009): 12–27.
- Beardsley, Timothy M. "Notes on Changing Practices in Data Publication." *BioScience* 65, no. 7 (2015): 645–50.
- Berman, Francine, and Vint Cerf. "Who Will Pay for Public Access to Research Data?" *Science* 341, no. 6146 (2013): 616–17.
- Berman, Helen M. "The Protein Data Bank: A Historical Perspective." *Acta Crystallographica Section A: Foundations of Crystallography* 64, no. 1 (2008): 88–95.
- Biosharing.org homepage. Accessed March 3, 2016. <http://biosharing.org>.
- Bishoff, Carolyn, and Lisa R. Johnston. "Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University." *Journal of Librarianship and Scholarly Communication* 3, no. 2 (2015): eP1231. doi:10.7710/2162-3309.1231.



- Borgman, Christine L. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press, 2015.
- . “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63, no. 6 (2012): 1059–78.
- Brook, Ryan K., and Stéphane M. McLachlan. “Trends and Prospects for Local Knowledge in Ecological and Conservation Research and Monitoring.” *Biodiversity and Conservation* 17, no. 14 (2008): 3501–12.
- CHORUS (Clearinghouse for the Open Research of the United States) homepage. Accessed June 15, 2016. <http://www.chorusaccess.org>.
- Choudhury, G. Sayeed. “Case Study in Data Curation at Johns Hopkins University.” *Library Trends* 57, no. 2 (Fall 2008): 211–20.
- COAR (Confederation of Open Access Repositories) homepage. Accessed June 15, 2016. <https://www.coar-repositories.org>.
- CODATA (Committee on Data for Science and Technology) homepage. Accessed May 17, 2016, <http://www.codata.org>.
- Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. Recommended Practice, CCSDS 650.0-M-2, Magenta Book, Issue 2. Washington DC: CCSDS Secretariat, June 2012. <http://public.ccsds.org/publications/archive/650x0m2.pdf>.
- COPDESS (Coalition on Publishing Data in the Earth and Space Sciences) homepage. Accessed June 15, 2016. <http://www.copdess.org>.
- Curd, Constanze, and Dirk Hoffmeister. “Research Data Management Services for a Multidisciplinary, Collaborative Research Project.” *Program: Electronic Library and Information Systems* 49, no. 4 (2015): 494–512, doi:10.1108/PROG-02-2015-0016.
- Darwin Core Task Group, “Darwin Core,” Biodiversity Information Standards (TDWG), last modified June 5, 2015. Accessed July 15, 2016. <http://rs.tdwg.org/dwcl/>.
- Data Foundation and Terminology Working Group, Research Data Alliance. Accessed May 18, 2016. <https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html>.
- Downs, Robert R., Ruth Duerr, Denise Hills, and Hampapuram K. Ramapriyan. “Data Stewardship in the Earth Sciences.” *D-Lib Magazine* 21, no. 7/8 (July/August 2015). doi:10.1045/july2015-downs.
- Duerr, Ruth. “Data Archives and Repositories.” In *Encyclopedia of Remote Sensing*. Edited by Eni G. Nijoku, 127–31. New York: Springer, 2014.
- EarthCube. “BCube.” Accessed August 3, 2016. <http://earthcube.org/group/bcube>.
- Earthdata, EOSDIS (Earth Observing System Data and Information System), NASA. Accessed May 18, 2016. <https://earthdata.nasa.gov/>.
- Fienberg, Stephen E., Margaret E. Martin, and Miron L. Straf, eds. *Sharing Research Data*. Washington, DC: National Academies Press, 1985. doi:10.17226/2033.
- Figshare homepage. Accessed May 18, 2016. <https://figshare.com>.
- Fox, Peter, and Ray Harris. “ICSU and the Challenges of Data and Information Management for International Science.” *Data Science Journal* 12 (2013): WDS1–WDS12. doi:10.2481/dsj.WDS-001.
- Gearheard, Shari. *When the Weather Is Uggianaqtug: Inuit Observations of Environmental Changes*, version 1, data set NSIDC 0650. Boulder, CO: NSIDC: National Snow and Ice Data Center, 2004). Accessed March 3, 2016, <https://nsidc.org/data/nsidc-0650>.
- Gearheard, Shari, Matthew Pocerlich, Ronald Stewart, Joelle Sanguya, and Henry P. Huntington. “Linking Inuit Knowledge and Meteorological Station Observations

- to Understand Changing Wind Patterns at Clyde River, Nunavut." *Climatic Change* 100, no. 2 (2010): 267–94.
- Gewin, Virginia. "Data Sharing: An Open Mind on Open Data." *Nature* 529, no. 7584 (2016): 117–19.
- Graham, Susan L., Marc Snir, and Cynthia A. Patterson. *Getting up to Speed: The Future of Supercomputing*. Washington, DC: National Academies Press, 2004. <http://www.nap.edu/catalog/11148/getting-up-to-speed-the-future-of-supercomputing>.
- Greenberg, Jane, Hollie C. White, Sarah Carrier, and Ryan Scherle. "A Metadata Best Practice for a Scientific Data Repository." *Journal of Library Metadata* 9, no. 3–4 (2009): 194–212.
- Gutmann, Myron, Kevin Schürer, Darrell Donakowski, and Hilary Beedham. "The Selection, Appraisal, and Retention of Social Science Data." *Data Science Journal* 3 (2006): 209–21. doi:10.2481/dsj.3.209.
- Harrison, Teresa M., Theresa A. Pardo, and Meghan Cook. "Creating Open Government Ecosystems: A Research and Development Agenda." *Future Internet* 4, no. 4 (2012): 900–28.
- Heidorn, P. Bryan. "Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends* 57, no. 2 (2008): 280–99.
- Henderson, Austen, and Morten Kyng. "There's No Place Like Home: Continuing Design in Use." In *Design at Work: Cooperative Design of Computer Systems*, edited by Joan Greenbaum and Morten Kyng, 219–40. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.
- Herzmann, Daryl E., Lori J. Abendroth, and Landon D. Bunderson. "Data Management Approach to Multidisciplinary Agricultural Research and Syntheses." *Journal of Soil and Water Conservation* 69, no. 6 (2014): 180A–185A.
- Hey, Tony, Stewart Tansley, and Kristin Tolle, eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, vol. 1. Redmond, WA: Microsoft Research, 2009. <http://research.microsoft.com/en-us/collaboration/fourthparadigm/contents.aspx>.
- Holdren, John P. "Increasing Access to the Results of Federally Funded Scientific Research." Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013. [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
- Hruby, Jill M., Dawn K. Manley, Ronald E. Stoltz, Erik K. Webb, and Joan B. Woodward. "The Evolution of Federally Funded Research and Development Centers." *Public Interest Report* 64, no. 1 (Spring 2011): 24–31. <http://www.fas.org/pubs/pir/2011spring/FFRDCs.pdf>.
- Interdisciplinary Earth Data Alliance homepage. Accessed May 18, 2016. <http://www.iedadata.org>.
- International Council for Science. *Ad Hoc Strategic Committee on Information and Data. Final Report to the ICSU Committee on Scientific Planning and Review*. Paris: International Council for Science, 2008. <http://www.icsu.org/publications/reports-and-reviews/scid-report>.
- International Council of Scientific Unions Panel on World Data Centres. *Guide to the World Data Center System*. Paris: International Council of Scientific Unions, 1996. <http://www.ukssdc.ac.uk/wdc/guide/wdcguide.pdf>.



- Inter-university Consortium for Political and Social Research. "Sustaining Domain Repositories for Digital Data: A Call for Change from an Interdisciplinary Working Group of Domain Repositories." June 24–25, 2013. <http://www.icpsr.umich.edu/icpsrweb/ICPSR/support/announcements/2013/09/sustaining-domain-repositories-for>.
- Jaguszewski, Janice M., and Karen Williams. *New Roles for New Times: Transforming Liaison Roles in Research Libraries*. Washington, DC: Association of Research Libraries, August 2013. <http://www.arl.org/component/content/article/6/2893>.
- Jisc. "List of Metadata Standards." Joint Information Systems Committee, Digital Curation Center. Accessed March 3, 2016. <http://www.dcc.ac.uk/resources/metadata-standards/list>.
- Jones, Sarah, Graham Pryor, and Angus Whyte. *How to Develop Research Data Management Services: A Guide for HEIs*. DCC How-to Guides. Edinburgh, UK: Digital Curation Center, 2013. <http://hdl.handle.net/11329/250>.
- Karasti, Helena, Karen S. Baker, and Florence Millerand. "Infrastructure Time: Long-Term Matters in Collaborative Development." *Computer Supported Cooperative Work (CSCW)* 19, no. 3–4 (2010): 377–415.
- Kuny, Terry. "The Digital Dark Ages? Challenges in the Preservation of Electronic Information." *International Preservation News*, no. 17 (1998): 8–13.
- Linguistic Data Consortium homepage. Accessed May 18, 2016. <https://www.ldc.upenn.edu>.
- Lynch, Clifford A. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." *ARL Bimonthly Report* 226 (February 2003): 1–7. <http://www.cni.org/publications/cliffs-pubs/institutional-repositories-infrastructure-for-scholarship/>.
- Lyon, Liz. *Dealing with Data: Roles, Rights, Responsibilities and Relationships*. Consultancy Report. University of Bath, UK: UKOLN, 2007. [http://opus.bath.ac.uk/412/1/dealing\\_with\\_data\\_report%2Dfinal.pdf](http://opus.bath.ac.uk/412/1/dealing_with_data_report%2Dfinal.pdf).
- Mauro, Francesco, and Preston D. Hardison. "Traditional Knowledge of Indigenous and Local Communities: International Debate and Policy Initiatives." *Ecological Applications* 10, no. 5 (2000): 1263–69.
- Mayernik, Matthew S. "Research Data and Metadata Curation as Institutional Issues." *Journal of the Association for Information Science and Technology* 67, no. 4 (April 2016): 973–93. doi:10.1002/asi.23425.
- Mayernik, Matthew S., Tim DiLauro, Ruth Duerr, Elliot Metsger, Anne E. Thessen, and G. Sayeed Choudhury. "Data Conservancy Provenance, Context, and Lineage Services: Key Components for Data Preservation and Curation." *Data Science Journal* 12 (2013): 158–71. doi:10.2481/dsj.12-039.
- McWhirter, Jeff, Christopher J. Crosby, Peter C. Griffith, Siri Khalsa, Matthew A. Lazzara, and W. Jeff Weber. "Rich Support for Heterogeneous Polar Data in RAMADDA." In *AGU Fall Meeting Abstracts*, vol. 1, 1632. Washington, DC: American Geophysical Union, 2013. <http://www.unidata.ucar.edu/committees/stratcom/2008spring/status-reports/ramadda.html>.
- Michener, William K., John Porter, Mark Servilla, and Kristin Vanderbilt. "Long Term Ecological Research and Information Management." *Ecological Informatics* 6, no. 1 (January 2011): 13–24.
- NASA Earthdata. "NASA Earth Science Data Preservation Content Specification." Accessed June 15, 2016. <https://earthdata.nasa.gov/standards/preservation-content-spec>.

- National Academy of Sciences. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington DC: National Academies Press, 2009. doi:10.17226/12615.
- National Research Council. *Government Data Centers: Meeting Increasing Demands*. Washington, DC: National Academies Press, 2003. doi:10.17226/10664.
- . *Preparing the Workforce for Digital Curation*. Washington, DC: National Academies Press, 2015. doi:10.17226/18590.
- . *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*. Washington, DC: National Academies Press, 1995. doi:10.17226/4871.
- . *Review of NASA's Distributed Active Archive Centers*. Washington, DC: National Academies Press, 1999. doi:10.17226/6396.
- National Science Board. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*. Arlington, VA: National Science Foundation, September 2005.
- Nativi, Stefano, Max Craglia, and Jay Pearlman. "Earth Science Infrastructures Interoperability: The Brokering Approach." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6, no. 3 (2013): 1118–29.
- Newton, Mark P., Christopher C. Miller, and Marianne Stowell Bracke. "Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force." *Collection Management* 36, no. 1 (2010): 53–67.
- NSF EarthCube-sponsored Roadmaps. Accessed May 18, 2016. <http://earthcube.org/type-document/roadmaps>.
- OpenICPSR homepage. Accessed May 18, 2016. <https://www.openicpsr.org>.
- Pampel, Heinz, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf. "Making Research Data Repositories Visible: The re3data.org Registry." *PLOS ONE* 8, no. 11 (2013): e78080. doi:10.1371/journal.pone.0078080.
- Parsons, Mark A., and Francine Berman. "The Research Data Alliance: Implementing the Technology, Practice and Connections of a Data Infrastructure." *Bulletin of the American Society for Information Science and Technology* 39, no. 6 (2013): 33–36.
- Parsons, Mark A., and Ruth Duerr. "Designating User Communities for Scientific Data: Challenges and Solutions." *Data Science Journal*, no. 4 (2006): 31–38. doi:10.2481/dsj.4.31.
- Parsons, Mark A., Oystein Godøy, Ellsworth LeDrew, Taco F. De Bruin, Bruno Danis, Scott Tomlinson, and David Carlson. "A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science." *Journal of Information Science* 37, no. 6 (2011): 555–69.
- Pejša, Stanislav, Shirley J. Dyke, and Thomas J. Hacker. "Building Infrastructure for Preservation and Publication of Earthquake Engineering Research Data." *International Journal of Digital Curation* 9, no. 2 (2014): 83–97.
- Pollock, Rufus. "Building the (Open) Data Ecosystem." *Open Knowledge Foundation Blog*, March 31, 2011. Accessed July 15, 2016. <http://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem>.
- Pulsifer, Peter, Shari Gearheard, Henry P. Huntington, Mark A. Parsons, Christopher McNeave, and Heidi S. McCann. "The Role of Data Management in Engaging Communities in Arctic Research: Overview of the Exchange for Local Observations and Knowledge of the Arctic (ELOKA)." *Polar Geography* 35, no. 3–4 (2012): 271–90.

- Ramapriyan, H. K. "The Role and Evolution of NASA's Earth Science Data Systems." Slides for presentation, IEEE EDS/CAS Chapter meeting, Camarillo, CA, August 19, 2015. Accessed February 13, 2016. <http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150018076.pdf>.
- Re3data (registry of Research Data Repositories) homepage. Accessed May 18, 2016. <http://re3data.org>.
- Research Information Network and the British Library. *Patterns of Information Use and Exchange: Case Studies of Researchers in the Life Sciences*. London: Research Information Network, November 2009. Accessed July 15, 2016. [http://www.rin.ac.uk/system/files/attachments/Patterns\\_information\\_use-REPORT\\_Nov09.pdf](http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf).
- Resnik, David B., and Gregg E. Dinse. "Do US Research Institutions Meet or Exceed Federal Requirements for Instruction in Responsible Conduct of Research? A National Survey." *Academic Medicine* 87, no. 9 (2012): 1237.
- Riley, Jenn. "Seeing Standards: A Visualization of the Metadata Universe." Poster, 2010. Accessed December 20, 2016. <http://jennriley.com/metadatamap/>.
- Rolando, Lizzy, Jake Carlson, Patricia Hswe, Susan Wells Parham, Brian Westra, and Amanda L. Whitmire. "Data Management Plans as a Research Tool." *Bulletin of the American Society for Information Science and Technology* 41, no. 5 (2015): 43–45.
- Royal Society Science Policy Centre. *Science as an Open Enterprise*. Final report. London: The Royal Society, June 2012. <https://royalsociety.org/-/media/policy/projects/sape/2012-06-20-saoc.pdf>.
- Salo, Dorothea. "Innkeeper at the Roach Motel." *Library Trends* 57, no. 2 (Fall 2008): 98–123.
- "Sayeed Choudhury on Data Stack Model," YouTube video, 5:53, posted by CLIRDLE, August 7, 2012. Accessed July 15, 2016. <https://www.youtube.com/watch?v=3MD-7KjZF34Y>.
- Shreeves, Sarah L., and Melissa H. Cragin. "Introduction: Institutional Repositories: Current State and Future." *Library Trends* 57, no. 2 (2008): 89–97.
- Starr, Joan, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, et al. "Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications." *PeerJ Computer Science* 1 (May 27, 2015): e1. doi:10.7717/peerj-cs.1.
- Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. "Data Sharing by Scientists: Practices and Perceptions." *PLOS ONE* 6, no. 6 (2011): e21101. doi:10.1371/journal.pone.0021101.
- Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide." *PLOS ONE* 10, no. 8 (2015): e0134826. doi:10.1371/journal.pone.0134826.
- US Geoscience Information Network Commons (GI-cat). Accessed August 3, 2016. <http://lab.usgin.org/applications/gi-cat>.
- Vardigan, Mary. "The DDI Matures: 1997 to the Present." *IASSIST Quarterly* 37, no. 1–4 (Spring 2013): 45–50.
- Walters, Tyler O. "The New Academic Library: Building Institutional Repositories to Support Changing Scholarly and Research Processes." In *Proceedings of the ACRL Thirteenth National Conference*, edited by Hugh A. Thompson, 56–63. Chicago:

- American Library Association, 2007. <http://www.ala.org/acrl/files/conferences/conf-sandpreconfs/national/baltimore/papers/56.pdf>.
- Walters, Tyler, and Judy Ruttenberg. "Shared Access Research Ecosystem." *Educause Review* 49, no. 2 (2014): 56–57.
- Whitmire, Amanda, Kristin Briney, Amy Nurnberger, Margaret Henderson, Thea Atwood, Margaret Janz, Wendy Kozlowski, Sherry Lake, Micah Vandegrift, and Lisa Zilinski. "A Table Summarizing the Federal Public Access Policies Resulting from the US Office of Science and Technology Policy Memorandum of February 2013," v. 5, April 18, 2016, Figshare, doi:10.6084/m9.figshare.1372041.
- Witt, Michael. "Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service." *Journal of Library Administration* 52, no. 2 (2012): 172–88.
- . "Institutional Repositories and Research Data Curation in a Distributed Environment." *Library Trends* 57, no. 2 (2008): 191–201.
- Witt, Michael, Jacob Carlson, D. Scott Brandt, and Melissa H. Cragin. "Constructing Data Curation Profiles." *International Journal of Digital Curation* 4, no. 3 (2009): 93–103.
- Worldwide Protein Data Bank homepage. Accessed May 18, 2016. <http://www ww p d b . o r g .>
- Yarmey, Lynn, and Karen S. Baker. "Towards Standardization: A Participatory Framework for Scientific Standard-Making." *International Journal of Digital Curation* 8, no. 1 (2013): 157–72.
- Young, Josh, W. Christopher Lenhardt, Mark Parsons, and Karl Benedict. "Taking Another Look at the Data Management Life Cycle: Deconstruction, Agile, and Community." Poster. In *AGU Fall 2014 Meeting Abstracts*, vol. 1, 3779. Washington, DC: American Geophysical Union, 2014. Accessed July 15, 2016. <http://abstractsearch.agu.org/meetings/2014/FM/IN51B-3779.html>.



## CHAPTER 2\*

# Institutional, Funder, and Journal Data Policies

*Kristin Briney, Abigail Goben, and Lisa Zilinski*

Data curation exists within a larger framework of laws and policies covering topics like copyright and data retention. These obligations must be considered in order to properly care for data as it is being created and preserved. While laws may transition slowly, the policies applying to research data by funding bodies, institutions, and journals have seen significant change since the turn of the century. These policies have directly impacted the practices of researchers and prompted the creation of data curation services by many libraries in partnership with their larger institutions.

This chapter examines three important categories of policies, primarily covered from the US perspective, that affect data curation practices in libraries: funding agency policies, institutional data policies, and journal data policies. While data professionals may be more familiar with funder and journal policies, institutional data policies are emerging as equally prevalent. Also, researchers across disciplines may encounter policies at a more granular level, such as for a specific research project or group, but these policies are less standardized and are therefore not covered in detail here.

Data policies are presently developing as researchers, institutions, funders, and journals look to improve research data management and sharing practices. As a result, standards for data policies have not yet been fully established. Potential topics covered in data policies include statements of data ownership, sharing

---

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

requirements, expected retention periods, access rights, and security issues. These may appear in a stand-alone policy or in multiple policy documents depending on the policy creator. While some homogenization may develop over time, the high levels of variance between policies from different sources—funders, institutions, and journals—and even between policies from similar sources, prevent the identification of consistent policy standards that cross all disciplinary and local boundaries.

Instead, this chapter outlines the similarities and differences between the general trends in funder, institutional, and journal policies, which are critical to understand. In particular, we must understand how the inconsistencies between these three policy types can cause challenges for researchers trying to meet overlapping requirements. This chapter will briefly recap the current state of these policy three areas, identify common overlap and variances, and suggest how we, as we undertake data curation, can navigate and influence this policy landscape.

## Funding Agency Data Policies

Funding agency policies have served a critical role in driving efforts on data curation as these policies primarily require researchers to preserve and share their data. While the policies themselves are mainly researcher-focused, libraries have an important role to play in this area due to their preservation expertise.

One of the first data policies by a major funding agency in the United States came from the National Institutes of Health (NIH) in 2003 and required researchers applying for direct annual costs of \$500,000 or more to create a plan for sharing their research data.<sup>1</sup> While this policy applied to a very limited number of grants awarded by the NIH, not including most R01 grants,<sup>2</sup> it was a clear indication that data is an important product of research that must be cared for, shared, and curated. Yet, the 2008 NIH Public Access Policy, which applied mainly to research articles, did not expand upon data as a research object to be shared.<sup>3</sup>

Then in 2011, the National Science Foundation (NSF) followed the NIH in adopting a data policy. This policy directed that all grant applications include a two-page-maximum data management plan (DMP) describing how the researchers would maintain, preserve, and make their data available.<sup>4</sup> The NSF specified that this supplemental documentation must include the types of data and other materials collected, applicable standards, provisions for sharing and providing access to the data for reuse, and plans for archiving the data.<sup>5</sup> More immediately impactful than the NIH policy, this policy meant that NSF grants with poor DMPs could be rejected, although the policy did not specify follow-up procedures for directorates to ensure compliance. Although the general policy applies

across the entire National Science Foundation, different divisions and directorates within the NSF could each provide more extensive policies and guidance for their individual programs. For example, the NSF Engineering Directorate required DMPs to specify the period of data retention, with a minimum requirement of three years,<sup>6</sup> and the Geological Sciences Directorate Division of Ocean Sciences stated that researchers must submit their data to an appropriate data center no later than two years after data collection.<sup>7</sup> The NSF policy was the inducement for many libraries to begin creating data services, not only around consulting on data management plans<sup>8</sup> but also around directly curating research data to satisfy both the data preservation and sharing portions of a DMP.

The NIH and NSF policies, while applying to a considerable number of researchers, were not systemic to the US federal funding system. That change came in 2013 when the White House Office of Science and Technology Policy (OSTP) published a memorandum on public access.<sup>9</sup> The OSTP memo covered not only public access to publications based on government-funded research, but also directed agencies with over \$100 million in annual research and development expenditures to require data management plans and maximize access to data from funded projects. Further, a White House Executive Order issued later in Spring 2013 required agencies to release their agency-generated data freely and in a machine-readable format, expanding the federal commitment to open and shared data.<sup>10</sup> As of early 2016, many of the covered funding agencies have enacted new requirements in response to the OSTP memo while others have only preliminary plans for compliance.

Requirements for data management plans and data curation and sharing are not limited to the United States. The 2007 OECD “Principles and Guidelines for Access to Research Data from Public Funding” was instrumental in bringing together thirty countries under the goal of improved access to research data.<sup>11</sup> Since then, significant work has been done, such as the Horizon 2020 program out of the European Commission,<sup>12</sup> and additional examples from the United Kingdom and Canada highlighted here.

The Research Councils United Kingdom (RCUK) and Wellcome Trust in the United Kingdom have enacted several data requirements.<sup>13</sup> These policies encourage researchers to make their data openly available as quickly as possible with a minimum number of restrictions. Similar to the US National Science Foundation, individual councils under the RCUK are also issuing their own data policies. The Engineering and Physical Sciences Research Council (EPSRC) is particularly notable in that the policy places heavy responsibility on the research organization—not just the researcher—for compliance. The policy dictates that organizations must make data openly available for a minimum of ten years with effective data curation across the data life cycle.<sup>14</sup> A more complete list of UK funder and institutional policies is available from the Digital Curation Centre.<sup>15</sup>



Canada is also developing data management policies for federally funded research. In 2015, three major Canadian research agencies—the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC)—put out a draft statement on principles of data management.<sup>16</sup> This draft policy builds on the Canadian government’s “Action Plan on Open Government,” which supports maximizing access to federally funded research and echoes funder policies from other countries by calling for data management plans and open data sharing.<sup>17</sup> The draft policy notably establishes the different responsibilities of researchers, research communities, institutions, and funders.

Beyond federal governments, an emerging trend among nonprofit funders is toward the requirement for data management plans and data preservation and sharing. Private nonprofit funding agencies, such as the Bill and Melinda Gates Foundation, are adopting such mandates.<sup>18</sup> The Gates Foundation policy was seen as an especially strong funder policy when it was announced in 2014 as it required immediate and open access to all data from all funded grants.<sup>19</sup> The following year, the Ford Foundation adopted a policy requiring all data from its sponsored grants be made available with a Creative Commons Attribution License (CC BY 4.0),<sup>20</sup> demonstrating funder interest not only in data sharing but also in allowing reuse and attribution. A major benefit of these data-sharing policies is that they require researchers to focus on better curation and management practices throughout the research process, knowing the data must be released at the end of a project.

For libraries engaged in grant writing and research, the US Institute of Museum and Library Services (IMLS) requires data sharing. The general guidelines for grants issued after December 2014 state, “If you collect and analyze data as part of an IMLS funded project, IMLS expects you to deposit data resulting from IMLS-funded research in a broadly accessible repository that allows the public to use the data without charge no later than the date upon which you submit your final report to IMLS. You should deposit the data in a machine-readable, non-proprietary digital format to maximize search, retrieval, and analysis.”<sup>21</sup>

The impetus behind funding agencies developing research data policies varies.<sup>22</sup> Altruistically, the goal is to expand access to research and increase the speed and replicability of science. Another argument is to allow taxpayers access to the research that they have funded. Additionally, facing increasing budget constraints, the agencies are focused on avoiding duplicative research and gaining a full return on their funding investment through data reuse in other projects. Funding agencies also may be looking to expand the possibility of their funded research being commercialized, available to the developing world or outside of academia, and improving education.

# Institutional Data Policies

With the increasing focus on data in the research and funding processes, individual academic institutions are creating and clarifying policies that outline data governance for their associated researchers. While many of these policies are more broadly concerned with intellectual property—a historic interest for universities with research resulting in patents—more research universities are starting to create stand-alone data policies. The 2013 ACRL SPEC Kit on research data management provides several examples of institutional data policies,<sup>23</sup> and a more recent review of 206 major research universities in the United States found that almost half had some policy covering research data—either an IP (15%) or a stand-alone data policy (29%).<sup>24</sup>

In contrast with funding agency data policies, university policies are often concerned with data ownership, retention, and access.<sup>25</sup> For example, many policies describe what should happen to the research data when the researcher leaves the institution and who is allowed access to this data in the meantime. Data ownership, when explicit in the policy, is often given to the university; this is likely a by-product of the funding system in the United States, where grants are given to the university to administer (with subsequent university compliance requirements) instead of to the researcher directly.

Institutional policies are not yet universal, and there is often discrepancy between existing institutional policies, which may exceed the differences observed between funder policies. While some policies are clear and comprehensive, others may impede the ability for researchers to conduct research and collaborate with their peers.

Exemplar institutional data policies should cover research data ownership, stewardship, and expectations as well as provide clear definitions, identify access and ownership claims to the data, specify retention periods, and lay out the responsibilities of all data stakeholders (including what happens if a researcher leaves the institution). Due to local differences, the ideal policy contents will vary between institutions and countries.<sup>26</sup>

There are several institutional policies that we recommend for review: the University of New Hampshire, the University of Minnesota, and the University of Massachusetts. These policies feature clear, explicit, and thorough language about what researchers should and should not do with their data. For example, the University of New Hampshire's "Policy on Ownership, Management, and Sharing of Research Data" provides straightforward definitions for investigators, research, research data, ownership, custodianship, and stewardship.<sup>27</sup> It acknowledges the authority of the investigators to do their own research, provides clear inclusion and exclusion of what constitutes research data, and defines roles and authority between the university administration and the investigator. Likewise, the "Research Data Management: Archiving, Ownership, Retention, Security,

Storage, and Transfer Policy” at the University of Minnesota is an example of direct writing.<sup>28</sup> The policy provides details on ownership and stewardship, data retention and archiving, research data transfer, researcher obligations, and data security. Specifically, this policy defines the role of the university libraries under the extensive responsibilities section with a number of specific examples. The “Policy on Data Ownership, Retention, and Access” at the University of Massachusetts Amherst also provides detailed definitions and covers data ownership, custody, quality, retention, and access.<sup>29</sup> Of particular note is the statement “When a collaboration comes to an end, and data was created during the collaboration, each member of the collaboration shall retain access to that data.”<sup>30</sup>

More general guidance on developing a research data management policy is provided by the Association of Southeastern Research Libraries in collaboration with the Southeastern Universities Research Association. The model policy is intended to be comprehensive, allowing institutions to select and adapt relevant sections as appropriate. The model includes suggested statements on the purpose of the policy, data ownership, stakeholders and their responsibilities, and potential related institutional policies.<sup>31</sup>

There are a variety of motivations for institutions to develop data policies. For example, universities have an interest in promoting and preserving the reputation of the institution and the researcher: where good data is known to be a product of the institution and its researchers, both entities can gain recognition for the data and research generated. Good policies may also prevent reputational damage when data is missing, lost, or found to be fraudulent. Another goal of an institutional data policy is to improve opportunities for commercialization, as controlling access to data and maintaining good data preservation and documentation are integral to patent applications. Finally, universities have a specific goal of data retention for educational reuse, as data is frequently shared between faculty and students in a “gift” culture that introduces students and early career researchers to the field.<sup>32</sup> Overall, however, institutional data policy is frequently focused on control of research data, which is sometimes at odds with mandates to curate this data for sharing with others.

## Journal Data Policies

Journal data policies add further complexity to the data policy landscape. These policies align with some of the recent changes to funding agency policies by pushing for greater access to research data. While still not ubiquitous in scholarly publishing, there are increasing journal and publisher requirements for researchers to make the supporting data available alongside the published journal article.

The actual journal requirements for data sharing fall on a spectrum from strict to loose. The *Public Library of Science* (PLOS) family of journals caused

controversy in 2014 for being one of the first large journals to strictly require data availability as a condition of publication.<sup>33</sup> Other journals, such as *Science* and *Nature*, expected researchers who published within their pages to provide data as requested but did not explicitly require data to be made openly available at the time of publication.<sup>34</sup> A further trend is data journals, where only the data with some supporting metadata is submitted for peer review.<sup>35</sup> We should be aware that journals in our own field are starting to enact similar expectations, such as for the *Journal of Librarianship and Scholarly Communication (JLSC)*.<sup>36</sup>

Beyond the basic expectation that data be made available, journals often recommend places for researchers to place their data to be in compliance with the policy. For journals with loose sharing expectations, it is often enough to simply provide access to the data when contacted rather than placing the data in a specific repository. For journals with strict data requirements, the journal may recommend a specific repository for data deposit, such as *JLSC*'s recommendation of its Dataverse instance,<sup>37</sup> or provide a list of recommended repositories across a variety of disciplines and subdisciplines.<sup>38</sup> Local institutional repositories run by libraries often do not appear in these directories or are listed with qualifications when they are.<sup>39</sup> Overall, journal policies reinforce the new data-sharing requirements of funder data policies and often take them a step further by specifying the preferred data repository for hosting.

Journals have their own motivations for enacting data policies. The principal incentive is to increase the reproducibility of the articles these journals publish. Greater scrutiny of research data can prevent the publication of problematic research and ensure that any subsequent retractions are easier to identify and resolve, both of which improve the quality and reputation of a journal. Open-access journals also have an altruistic motivation to expand their open mission into the data realm.

## Navigating the Data Policy Landscape for Curation

Libraries undertaking data curation must be aware of funding agency, institutional, and journal data policies as these policies can directly affect local curation practices. Part of this awareness requires the ability to navigate the variances that frequently exist between the policy types. Thankfully, there are also a few areas of policy agreement that can further strengthen curation efforts.

With respect to policy agreement, both funder and institutional policies often include a requirement about data retention after the end of project. This is a direct response to the fact that researchers often have difficulty with data

retention, with Vines and colleagues finding that research data availability falls by approximately 17 percent per year after the paper is published due to the data becoming “either lost or on inaccessible storage.”<sup>40</sup> Having a mandated policy on retention provides leverage when working with researchers, who often think of retention in terms of long-term storage instead of involving the preservation actions necessary to make sure that the data remains usable in the future.<sup>41</sup> By relying on the policies, we can ensure that data remains not only available but usable well after a project is complete.

However, while funder and institutional data policies often include retention mandates, retention times can sometimes conflict. The minimum retention period for data from government-funded research, per the US Office of Management and Budget (OMB) Uniform Guidance, is three years after the completion of the grant.<sup>42</sup> Where data retention times are stated in university policy, they can often be three, five, or seven years, or a fixed time may not be specified.<sup>43</sup> Retention periods may also vary by discipline. This creates confusion for researchers in how long they actually need to retain data and whose policy takes precedence. In practice, longer retention times are preferred, especially in light of a two recent retractions of six- and eight-year-old papers where the original data could not be located to address concerns about the research.<sup>44</sup> Retention is unfortunately more complicated for sensitive data; in this case, it may be best to refer questions to the local institutional review board (IRB), the institution’s chief information officer, or similar IT representatives to determine local practice. In general, libraries should recommend that stated retention times be treated as minimums, with a preference for longer, but not indefinite, retention periods.

A second area of overlap between institutional and funder policies is that responsibility for the data often falls to both the researcher and the university. US funding agency policy places sharing and retention responsibility on the principal investigator (PI) of the grant in addition to mandating compliance measures from the university overall. Institutional policy, on the other hand, often designates the PI as the data steward who makes most of the decisions about the data while the university is the actual data owner. This further varies by institution and disciplinary practices. In general, the institution is held responsible for the compliance of its researchers and has a financial interest in meeting these requirements. In terms of data curation efforts, these shared responsibilities lend authority to libraries to preserve data on behalf of the university and its commitments, as libraries are a natural home for this type of work.

There is a downside to this overlap, as the university will not often exert its claim of data ownership under local policy unless extreme measures are involved. These measures can include researcher misconduct, avoiding sensitive data breaches and large-scale audits, and issues when prestigious research is in-

volved or where the university has a large financial stake in the research or research products, in addition to routine compliance requirements from funders. The 2015 court case between the University of California–San Diego (UCSD) and the University of Southern California (USC) illustrates such an example. UCSD sued USC and former UCSD researcher Paul Aisen for attempts to cut UCSD off from grant money and the longitudinal data from the Alzheimer’s Disease Cooperative Study when the PI, Aisen, tried to move the center and many of its researchers from UCSD to USC.<sup>45</sup> In this case, UCSD used its backing from the NIH, which awarded the grant to UCSD and wished to continue to do so, and its data ownership policy to block Aisen and USC from their attempts to transfer the research project. While many researchers have likely left UCSD in possession of their data and grant funding, the prestige and value of this research prompted UCSD to exert its claim to research via its data policy. As research funding becomes more competitive, such issues are likely to arise more frequently.

These ownership issues may be further complicated in the case of unfunded research, collaborative research, or research where there is not a sole primary investigator at one institution. Researchers may want to share their data but feel confused when policy is not clear about external collaborative data sharing but still requires institutional ownership of the data.

Journal data policies deviate from funder and institutional policy in this area in that they rarely identify institutions as having any role at all in policy compliance. This is evident by how infrequently institutional repositories show up in lists of recommended repositories and the qualifications upon them, such as minting DataCite DOIs and placing data in an external backup repository, when they do.<sup>46</sup> It is useful to be aware of these external requirements when developing repository services as well as actively promoting institutional repositories to journals as a way to satisfy their data-sharing requirements. An example of this promotion comes from a group of data librarians and curators called DataCure who in 2015 called upon PLOS to include institutional repositories as a recommended place for archiving data.<sup>47</sup>

Most often, funding agency, institutional, and journal data policies disagree because the three policy types have fundamentally different intents. Funding agencies are usually concerned with data management, preservation, and sharing as they seek to prevent duplication and improve return on investment. In contrast, institutional data policies are more focused on data ownership and data control as they seek to maintain reputation and commercial control of intellectual property. Journal data policies, on the other hand, aim to improve the reproducibility of the journal’s published articles by providing access to the corresponding data. All policy types aim to lengthen the life cycle of research data, but two do so by promoting openness and while the third does so by putting restrictions on the data.<sup>48</sup> While the OECD Principles, OSTP Memo, and

Canadian council policies demonstrate the emerging standardization of data policies across the major government funding agencies,<sup>49</sup> no similar motion has yet occurred for institutional and journal data policies. Therefore, libraries engaging in data curation have a role in developing institutional data policy where it does not exist and lobbying for the inclusion of local data repositories in current and future journal data policies.

Another challenge to curation is that the three policy types have different enforcement mechanisms. Funding agencies have more leverage here as they can withhold money from those institutions that do not comply. Universities seldom have this option for enforcement. Journals can either refuse to publish articles by noncompliant researchers or retract them later.<sup>50</sup> Additionally, many researchers may not think to look to their libraries for support, and libraries rarely have the authority to enforce improved data curation practices, which compounds these curation problems.<sup>51</sup> Libraries involved in data curation should consider other motivations for researcher participation in data curation besides direct compliance.

Finally, one of the biggest challenges comes from when policies are diametrically opposed. The question then becomes: which policy wins? There is no clear answer to this question at present, so local practice may vary as institutions continue to develop data curation policies and services. Libraries, however, already support researchers in evaluating journals for publishing and can apply that skill set here, holding a key position from which to identify where policies conflict and to collaborate with administrators, researchers, and journal editors to resolve the effects of disparate policies on data curation.

Navigating this shifting policy landscape can be a challenge for libraries working to curate research data. There are, however, many things that libraries can do in this area:

- Identify opportunities for the library to act on behalf of the institution and its obligations to preserve and retain data.
- Advocate locally that the library is a natural home for these tasks, which might not get accomplished without the library's leadership.
- Collaborate with institutional administrators to either develop or improve institutional data policies.
- Be proactive in advocating the library's role in compliance with journal editors.
- Leverage existing policies to promote services.
- Provide guidance to researchers on complying with (sometimes conflicting) policy requirements.

There is no one best way to navigate the changing policy landscape, but by being aware of the myriad requirements, libraries can use them to the best advantage.



## Summary

Libraries engaged with data curation must be knowledgeable about the funding agency, institutional, and journal data policies that influence researcher responsibilities. Awareness of these evolving policies will enhance the library services for research data curation. We also have the opportunity to influence development or modification of our institutional policies to improve local data curation practices. Future navigation of policies will be important until further clarity and harmonization are established between funding agencies, institutions, and journals.

## Notes

1. National Institutes of Health, "NIH Data Sharing Policy and Implementation Guidance," Last updated March 5, 2003, [http://grants.nih.gov/grants/policy/data\\_sharing/data\\_sharing\\_guidance.htm](http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm).
2. Sally Rockey, "2014 By the Numbers," *Rock Talk* (blog), NIH Extramural Nexus, December 31, 2014, <https://nexus.od.nih.gov/all/2014/12/31/2014-by-the-numbers/>.
3. National Institutes of Health, "NIH Public Access Policy Details," accessed October 27, 2015, <https://publicaccess.nih.gov/policy.htm>.
4. National Science Foundation, "Dissemination and Sharing of Research Results," November 30, 2010, <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>.
5. National Science Foundation, "Plans for Data Management and Sharing of the Products of Research," in chapter 2, "Proposal Preparation Instructions," *Proposal and Award Policies and Procedures Guide*, NSF 15-1, OMB 3145-0058 (Washington, DC: Office of Management and Budget, 2014), [http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg\\_2.jsp#dmp](http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp).
6. National Science Foundation Engineering Directorate, *Data Management for NSF Engineering Directorate Proposals and Awards* (Arlington, VA: National Science Foundation), accessed November 17, 2015, [http://nsf.gov/eng/general/ENG\\_DMP\\_Policy.pdf](http://nsf.gov/eng/general/ENG_DMP_Policy.pdf).
7. National Science Foundation, *Division of Ocean Sciences Sample and Data Policy*, NSF 11060 (Arlington, VA: National Science Foundation, May 2011), <http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf>.
8. Regina Raboin, Rebecca Reznik-Zellen, and Dorothea Salo, "Forging New Service Paths: Institutional Approaches to Providing Research Data Management Services," *Journal of eScience Librarianship* 1, no. 3 (2012), doi:10.7191/jeslib.2012.1021; Karen Antell, Jody Bales Foote, Jaymie Turner, and Brain Shults, "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions," *College and Research Libraries* 75, no. 4 (July 1, 2014): 557–74, doi:10.5860/crl.75.4.557.
9. John P. Holdren, "Increasing Access to the Results of Federally Funded Scientific Research," Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013, [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
10. Barack Obama, "Executive Order: Making Open and Machine Readable the New Default for Government Information," May 9, 2013, <http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->



11. Organisation for Economic Co-operation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding* (Paris: OECD Publishing, April 2007), <http://www.oecd.org/sti/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm>.
12. European Commission, Horizon 2020 homepage, 2014, <https://ec.europa.eu/programmes/horizon2020/>.
13. Research Councils UK, “RCUK Common Principles on Data Policy,” accessed November 1, 2015, <http://www.rcuk.ac.uk/research/datapolicy/>; Wellcome Trust, “Policy on Data Management and Sharing,” accessed October 21, 2014, <http://www.wellcome.ac.uk/about-us/policy/policy-and-position-statements/wtx035043.htm>.
14. Engineering and Physical Sciences Research Council, “EPSRC Policy Framework on Research Data,” March 2011, <https://www.epsrc.ac.uk/about/standards/researchdata/>.
15. Digital Curation Centre, “Overview of Funders’ Data Policies,” accessed October 21, 2014, <http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies>; L. Horton and Digital Curation Centre, “Overview of UK Institution RDM Policies,” Digital Curation Centre, 2014, <http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies>.
16. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, “Draft Tri-Agency Statement of Principles on Digital Data Management,” July 9, 2015, last modified July 20, 2015, <http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1>.
17. “Canada’s Action Plan on Open Government 2014–16,” chapter IV-B, “Open Data—Open Data without Borders,” Open Government website, accessed February 16, 2016, <http://open.canada.ca/en/content/canadas-action-plan-open-government-2014-16#ch4-2>.
18. Bill and Melinda Gates Foundation, “Bill & Melinda Gates Foundation Open Access Policy,” accessed January 5, 2015, <http://www.gatesfoundation.org/how-we-work/general-information/open-access-policy>.
19. Richard Van Noorden, “Gates Foundation Announces World’s Strongest Policy on Open Access Research,” *Nature News Blog*, November 21, 2014, <http://blogs.nature.com/news/2014/11/gates-foundation-announces-worlds-strongest-policy-on-open-access-research.html>.
20. Ford Foundation, “Ford Foundation Expands Creative Commons Licensing for All Grant-Funded Projects,” news release, February 3, 2015, <https://www.fordfoundation.org/the-latest/news/ford-foundation-expands-creative-commons-licensing-for-all-grant-funded-projects/>.
21. Institute for Museum and Library Services, *General Terms and Conditions for IMLS Discretionary Grant and Cooperative Agreement Awards: For Awards Made After December 26, 2014* (Washington, DC: Institute for Museum and Library Services, March 30, 2015), 15, [https://www.imls.gov/sites/default/files/gtc\\_afterdec2014\\_0315.pdf](https://www.imls.gov/sites/default/files/gtc_afterdec2014_0315.pdf).
22. Christine L. Borgman, “The Conundrum of Sharing Research Data,” *Journal of the American Society for Information Science and Technology* 63, no. 6 (June 2012): 1059–78, doi:10.1002/asi.22634.
23. David Fearon Jr., Betsy Gunia, Sherry Lake, Barbara E. Pralle and Andrew L. Sallans, *Research Data Management Services, SPEC Kit 334* (Washington, DC: Association of Research Libraries, July 2013), <http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/>.

24. Kristin Briney, Abigail Goblen, and Lisa Zilinski, "Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies," *Journal of Librarianship and Scholarly Communication* 3, no. 1 (Summer 2015): eP1232, doi:10.7710/2162-3309.1232.
25. Ibid.
26. Ibid.; Horton and Digital Curation Centre, "Overview of UK Institution RDM Policies."
27. University of New Hampshire, "C. UNH Policy on Ownership, Management, and Sharing of Research Data | University System of New Hampshire," USNH Online Policy Manual, accessed November 17, 2015, <https://www.usnh.edu/policy/unh/viii-research-policies/c-unh-policy-ownership-management-and-sharing-research-data>.
28. University of Minnesota, "Research Data Management: Archiving, Ownership, Retention, Security, Storage, and Transfer," Administrative Policy, accessed November 17, 2015, <http://policy.umn.edu/research/researchdata>.
29. University of Massachusetts Amherst, *Special Report of the Research Council Concerning the Policy on Data Ownership, Retention, and Access at the University of Massachusetts Amherst*, Sen. Doc. No. 06-047, May 18, 2006, [http://www.umass.edu/research/sites/default/files/policy\\_on\\_data\\_ownership\\_retention\\_and\\_access\\_0.pdf](http://www.umass.edu/research/sites/default/files/policy_on_data_ownership_retention_and_access_0.pdf).
30. Ibid., 4.
31. Nathan Hall, Bill Corey, Wendy Mann, Tom Wilson, and ASERL/SURA Research Data Coordinating Committee, *Model Language for Research Data Management Policies* (Durham, NC: Association of Southeastern Research Libraries, January 2013), [http://www.aserl.org/wp-content/uploads/2013/01/ASERL-SURA\\_Model\\_Language\\_RDM\\_Policy\\_Language\\_FINAL.pdf](http://www.aserl.org/wp-content/uploads/2013/01/ASERL-SURA_Model_Language_RDM_Policy_Language_FINAL.pdf).
32. Borgman, "The Conundrum of Sharing Research Data."
33. Liz Silva, "PLOS' New Data Policy: Public Access to Data," *EveryONE* (blog), February 24, 2014, <http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/>; Theo Bloom, "PLOS' New Data Policy: Part Two," *EveryONE* (blog), March 8, 2014, <http://blogs.plos.org/everyone/2014/03/08/plos-new-data-policy-public-access-data/>; Ivan Oransky, "Following Criticism, PLOS Apologizes, Clarifies New Data Policy," Retraction Watch website, March 9, 2014, <http://retractionwatch.com/2014/03/09/following-criticism-plos-apologizes-clarifies-new-data-policy/>; Carl Boettiger, "Plos Data Sharing Policy Reflections," May 30, 2014, <http://www.carlboettiger.info/2014/05/30/PLoS-data-sharing-policy-reflections.html>.
34. *Science*, "Science: Editorial Policies," July 2, 2015, <http://www.sciencemag.org/authors/science-editorial-policies>; Nature Publishing Group, "Availability of Data, Material, and Methods," Authors and Referees: Policies, November 18, 2014, <http://www.nature.com/authors/policies/availability.html#data>.
35. Katherine G. Akers, "A Growing List of Data Journals," *Data@MLibrary* (blog), May 9, 2014, <https://mlibrarydata.wordpress.com/2014/05/09/data-journals/>; Pauline Ward, "Sources of Dataset Peer Review," DataShare wiki, last modified January 12, 2016, <https://www.wiki.ed.ac.uk/display/dataset+peer+review>.
36. JLSC Editorial Board, "The Article Is Not Enough: Introducing the JLSC Data Sharing Policy," *Journal of Librarianship and Scholarly Communication* 2, no. 3 (2014), doi:10.7710/2162-3309.1186.
37. "Journal of Librarianship and Scholarly Communication Instructions for Authors," accessed February 16, 2016, <http://jpsc-pub.org/about/submissions/#data>.
38. BioMed Central, "Editorial Policies: Availability of Data and Materials," accessed Febru-

- ary 7, 2016, <https://www.biomedcentral.com/submissions/editorial-policies#availability+of+data+and+materials>; *PLOS ONE*, “Data Availability,” 2016, <http://journals.plos.org/plosone/s/data-availability>.
39. *Scientific Data*, “Recommended Repositories,” accessed February 7, 2016, <http://www.nature.com/sdata/data-policies/repositories>.
  40. Timothy H. Vines, Arianne Y. K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, et al., “The Availability of Research Data Declines Rapidly with Article Age,” *Current Biology* 24, no. 1 (January 2014): 95, doi:10.1016/j.cub.2013.11.014.
  41. Johns Hopkins University Data Management Services, “About Storing and Archiving Your Research Data,” accessed November 2, 2015, <https://dmp.data.jhu.edu/preserve-share-research-data/preserve-archive/>.
  42. Office of Management and Budget, “Circular A-110 Revised 11/19/93 As Further Amended 9/30/99,” White House website, last updated September 30, 1999, [http://www.whitehouse.gov/omb/circulars\\_a110](http://www.whitehouse.gov/omb/circulars_a110).
  43. University of New Hampshire, “C. UNH Policy on Ownership”; University of Wisconsin-Madison, “Policy on Data Stewardship, Access, and Retention,” accessed March 4, 2016, <https://kb.wisc.edu/gradsch/page.php?id=34404>; University of Iowa, “6g. Data Management: Research Records,” Researcher Handbook, accessed March 4, 2016, <http://researcherhandbook.research.uiowa.edu/6g-data-management-research-records>; University of Minnesota, “Research Data Management.”
  44. Ivan Oransky, “JCI Paper Retracted for Duplicated Panels after Authors Can’t Provide Original Data,” Retraction Watch website, July 19, 2013, [http://retractionwatch.com/2013/07/19/jci-paper-retracted-for-duplicated-panels-after-authors-cant-provide-original-data/?utm\\_source=dldr.it&utm\\_medium=twitter/](http://retractionwatch.com/2013/07/19/jci-paper-retracted-for-duplicated-panels-after-authors-cant-provide-original-data/?utm_source=dldr.it&utm_medium=twitter/); Ross Keith, “Author Objects to Retraction for Not ‘Faithfully Represented’ Immunology Figures,” Retraction Watch website, October 2, 2015, <http://retractionwatch.com/2015/10/02/author-objects-to-retraction-for-not-faithfully-represented-immunology-figures/>.
  45. Bradley J. Fikes, “UC San Diego Sues USC and Scientist, Alleging Conspiracy to Take Funding, Data,” *LA Times*, July 5, 2015, <http://www.latimes.com/local/education/lame-ucsd-lawsuit-20150706-story.html>; Scott Jaschik, “UCSD 1, USC 0,” *Inside Higher Ed*, July 27, 2015, <https://www.insidehighered.com/news/2015/07/27/ucsd-wins-key-round-legal-fight-usc-over-huge-research-project>.
  46. *Scientific Data*, “Recommended Repositories.”
  47. Stephen Abrams, Eugene Barsky, Kristin Briney, K. Jane Burpee, Jake Carlson, Heather Coates, Jennifer Doty, et al., “Open Letter to PLoS—Libraries Role in Data Curation,” Datacure, November 2015, <https://datacurepublic.wordpress.com/open-letter-to-plos-libraries-role-in-data-curation/>.
  48. Spencer D. C. Keralis, Shannon Stark, Martin Halbert, and William E. Moen, “Research Data Management in Policy and Practice: The DataRes Project,” In *Research Data Management: Principles, Practices, and Prospects*, CLIR Publication No. 160 (Washington, DC: Council on Library and Information Resources, 2013), 16.
  49. Organisation for Economic Co-operation, *OECD Principles and Guidelines*; Holdren, “Increasing Access to the Results of Federally Funded Scientific Research”; Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada., “Draft Tri-Agency Statement of Principles.”

50. Chelsey Coombs, "Neuroscience Paper Retracted after Colleagues Object to Data Publication," Retraction Watch website, December 31, 2015, <http://retractionwatch.com/2015/12/31/neuroscience-paper-retracted-after-colleagues-object-to-data-publication/>.
51. Raboin, Reznik-Zellen, and Salo, "Forging New Service Paths."

## Bibliography

- Abrams, Stephen, Eugene Barsky, Kristin Briney, K. Jane Burpee, Jake Carlson, Heather Coates, Jennifer Doty, et al. "Open Letter to PLoS—Libraries Role in Data Curation." DataCure website, November 2015. <https://datacurepublic.wordpress.com/open-letter-to-plos-libraries-role-in-data-curation/>.
- Akers, Katherine G. "A Growing List of Data Journals." *Data@MLibrary* (blog), May 9, 2014. <https://mlibrarydata.wordpress.com/2014/05/09/data-journals/>.
- Antell, Karen, Jody Bales Foote, Jaymie Turner, and Brian Shults. "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions." *College and Research Libraries* 75, no. 4 (July 2014): 557–74. doi:10.5860/crl.75.4.557.
- Bill and Melinda Gates Foundation. "Bill & Melinda Gates Foundation Open Access Policy." Accessed January 5, 2015. <http://www.gatesfoundation.org/how-we-work/general-information/open-access-policy>.
- BioMed Central. "Editorial Policies: Availability of Data and Materials." Accessed February 7, 2016. <https://www.biomedcentral.com/submissions/editorial-policies#availability+of+data+and+materials>.
- Bloom, Theo. "PLOS' New Data Policy: Part Two." *EveryONE* (blog). March 8, 2014. <http://blogs.plos.org/everyone/2014/03/08/plos-new-data-policy-public-access-data/>.
- Boettiger, Carl. "Plos Data Sharing Policy Reflections," May 30, 2014. <http://www.carlboettiger.info/2014/05/30/PLoS-data-sharing-policy-reflections.html>.
- Borgman, Christine L. "The Conundrum of Sharing Research Data." *Journal of the American Society for Information Science and Technology* 63, no. 6 (June 2012): 1059–78. doi:10.1002/asi.22634.
- Briney, Kristin, Abigail Goben, and Lisa Zilinski. "Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies." *Journal of Librarianship and Scholarly Communication* 3, no. 1 (Summer 2015): eP1232. doi:10.7710/2162-3309.1232.
- "Canada's Action Plan on Open Government 2014–16." Chapter IV-B, "Open Data—Open Data without Borders." Open Government website. Accessed February 16, 2016. <http://open.canada.ca/en/content/canadas-action-plan-open-government-2014-16#ch4-2>.
- Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada. "Draft Tri-Agency Statement of Principles on Digital Data Management," July 9, 2015. Last modified July 20, 2015. <http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1>.
- Coombs, Chelsey. "Neuroscience Paper Retracted after Colleagues Object to Data Publication." Retraction Watch website, December 31, 2015. <http://retractionwatch.com/2015/12/31/neuroscience-paper-retracted-after-colleagues-object-to-data-publication/>.

- com/2015/12/31/neuroscience-paper-retracted-after-colleagues-object-to-data-publication/.
- Digital Curation Centre. "Overview of Funders' Data Policies." Accessed October 21, 2014. <http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies>.
- Engineering and Physical Sciences Research Council. "EPSRC Policy Framework on Research Data," March 2011. <https://www.epsrc.ac.uk/about/standards/researchdata/>.
- European Commission. Horizon 2020 homepage. 2014. <https://ec.europa.eu/programmes/horizon2020/>.
- Fearon, David Jr., Betsy Gunia, Sherry Lake, Barbara E. Pralle, and Andrew L. Sallans. *Research Data Management Services, SPEC Kit 334*. Washington, DC: Association of Research Libraries, July 2013. <http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/>.
- Fikes, Bradley J. "UC San Diego Sues USC and Scientist, Alleging Conspiracy to Take Funding, Data." *LA Times*. July 5, 2015. <http://www.latimes.com/local/education/la-me-ucsd-lawsuit-20150706-story.html>.
- Ford Foundation. "Ford Foundation Expands Creative Commons Licensing for All Grant-Funded Projects." News release. February 3, 2015. <https://www.fordfoundation.org/the-latest/news/ford-foundation-expands-creative-commons-licensing-for-all-grant-funded-projects/>.
- Hall, Nathan, Bill Corey, Wendy Mann, Tom Wilson, and ASERL/SURA Research Data Coordinating Committee. *Model Language for Research Data Management Policies*. Durham, NC: Association of Southeastern Research Libraries, January 2013. [http://www.aserl.org/wp-content/uploads/2013/01/ASERL-SURA\\_Model\\_Language\\_RDM\\_Policy\\_Language\\_FINAL.pdf](http://www.aserl.org/wp-content/uploads/2013/01/ASERL-SURA_Model_Language_RDM_Policy_Language_FINAL.pdf).
- Holdren, John P. "Increasing Access to the Results of Federally Funded Scientific Research." Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013. [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
- Horton, L., and Digital Curation Centre. "Overview of UK Institution RDM Policies." Digital Curation Centre, 2014. <http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies>.
- Institute for Museum and Library Services. *General Terms and Conditions for IMLS Discretionary Grant and Cooperative Agreement Awards: For Awards Made After December 26, 2014*. Washington, DC: Institute for Museum and Library Services, March 30, 2015. [https://www.imls.gov/sites/default/files/gtc\\_afterdec2014\\_0315.pdf](https://www.imls.gov/sites/default/files/gtc_afterdec2014_0315.pdf).
- Jaschik, Scott. "UCSD 1, USC 0." *Inside Higher Ed*, July 27, 2015. <https://www.insidehighered.com/news/2015/07/27/ucsd-wins-key-round-legal-fight-usc-over-huge-research-project>.
- JLSC Editorial Board. "The Article Is Not Enough: Introducing the JLSC Data Sharing Policy." *Journal of Librarianship and Scholarly Communication* 2, no. 3 (2014). doi:10.7710/2162-3309.1186.
- Johns Hopkins University Data Management Services. "About Storing and Archiving Your Research Data." Accessed November 2, 2015. <https://dmp.data.jhu.edu/preserve-share-research-data/preserve-archive/>.
- "*Journal of Librarianship and Scholarly Communication* Instructions for Authors." Accessed February 16, 2016. <http://jls-public.org/about/submissions/#data>.

- Keith, Ross. "Author Objects to Retraction for Not 'Faithfully Represented' Immunology Figures." Retraction Watch website, October 2, 2015. <http://retractionwatch.com/2015/10/02/author-objects-to-retraction-for-not-faithfully-represented-immunology-figures/>.
- Keralis, Spencer D. C., Shannon Stark, Martin Halbert, and William E. Moen. "Research Data Management in Policy and Practice: The DataRes Project." In *Research Data Management: Principles, Practices, and Prospects*, 16–38. CLIR Publication No. 160. Washington, DC: Council on Library and Information Resources, 2013.
- National Institutes of Health. "NIH Data Sharing Policy and Implementation Guidance." Last updated March 5, 2003. [http://grants.nih.gov/grants/policy/data\\_sharing/data\\_sharing\\_guidance.htm](http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm).
- . "NIH Public Access Policy Details." Accessed October 27, 2015. <https://publicaccess.nih.gov/policy.htm>.
- National Science Foundation. "Dissemination and Sharing of Research Results." November 30, 2010. <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>.
- . *Division of Ocean Sciences Sample and Data Policy*. NSF 11060. Arlington, VA: National Science Foundation, May 2011. <http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf>.
- . "Plans for Data Management and Sharing of the Products of Research." In chapter 2, "Proposal Preparation Instructions," *Proposal and Award Policies and Procedures Guide*. NSF 15-1. OMB 3145-0058. Washington, DC: Office of Management and Budget, 2014. <http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg2.jsp#dmp>.
- National Science Foundation Engineering Directorate. *Data Management for NSF Engineering Directorate Proposals and Awards*. Accessed November 17, 2015. Arlington, VA: National Science Foundation. [http://nsf.gov/eng/general/ENG\\_DMP\\_Policy.pdf](http://nsf.gov/eng/general/ENG_DMP_Policy.pdf).
- Nature Publishing Group. "Availability of Data, Material, and Methods." Authors and Referees: Policies, November 18, 2014. <http://www.nature.com/authors/policies/availability.html#data>.
- Obama, Barack. "Executive Order: Making Open and Machine Readable the New Default for Government Information." May 9, 2013. <http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->
- Office of Management and Budget. "Circular A-110 Revised 11/19/93 As Further Amended 9/30/99." White House website. Last updated September 30, 1999. [http://www.whitehouse.gov/omb/circulars\\_a110](http://www.whitehouse.gov/omb/circulars_a110).
- Oransky, Ivan. "Following Criticism, PLOS Apologizes, Clarifies New Data Policy." Retraction Watch website, March 9, 2014. <http://retractionwatch.com/2014/03/09/following-criticism-plos-apologizes-clarifies-new-data-policy/>.
- . "JCI Paper Retracted for Duplicated Panels after Authors Can't Provide Original Data." Retraction Watch website, July 19, 2013. [http://retractionwatch.com/2013/07/19/jci-paper-retracted-for-duplicated-panels-after-authors-cant-provide-original-data/?utm\\_source=dlvr.it&utm\\_medium=twitter/](http://retractionwatch.com/2013/07/19/jci-paper-retracted-for-duplicated-panels-after-authors-cant-provide-original-data/?utm_source=dlvr.it&utm_medium=twitter/).
- Organisation for Economic Co-operation and Development. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: OECD Publishing, April 2007. <http://www.oecd.org/sti/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm>.

- PLOS ONE*. "Data Availability." 2016. <http://journals.plos.org/plosone/s/data-availability>.
- Raboin, Regina, Rebecca Reznik-Zellen, and Dorothea Salo. "Forging New Service Paths: Institutional Approaches to Providing Research Data Management Services." *Journal of eScience Librarianship* 1, no. 3 (2012). doi:10.7191/jeslib.2012.1021.
- Research Councils UK. "RCUK Common Principles on Data Policy." Accessed November 1, 2015. <http://www.rcuk.ac.uk/research/datapolicy/>.
- Rockey, Sally. "2014 By the Numbers." *Rock Talk* (blog), NIH Extramural Nexus, December 31, 2014. <https://nexus.od.nih.gov/all/2014/12/31/2014-by-the-numbers/>.
- Science*. "Science: Editorial Policies," July 2, 2015. <http://www.sciencemag.org/authors/science-editorial-policies>.
- Scientific Data*. "Recommended Repositories." Accessed February 7, 2016. <http://www.nature.com/sdata/data-policies/repositories>.
- Silva, Liz. "PLOS' New Data Policy: Public Access to Data." *EveryONE* (blog), February 24, 2014. <http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/>.
- University of Iowa. "6g. Data Management: Research Records." Researcher Handbook. Accessed March 4, 2016. <http://researcherhandbook.research.uiowa.edu/6g-data-management-research-records>.
- University of Massachusetts Amherst. *Special Report of the Research Council Concerning the Policy on Data Ownership, Retention, and Access at the University of Massachusetts Amherst*. Sen. Doc. No. 06-047. May 18, 2006, [http://www.umass.edu/research/sites/default/files/policy\\_on\\_data\\_ownership\\_retention\\_and\\_access\\_0.pdf](http://www.umass.edu/research/sites/default/files/policy_on_data_ownership_retention_and_access_0.pdf).
- University of Minnesota. "Research Data Management: Archiving, Ownership, Retention, Security, Storage, and Transfer." Administrative Policy. Accessed November 17, 2015. <http://policy.umn.edu/research/researchdata>.
- University of New Hampshire. "C. UNH Policy on Ownership, Management, and Sharing of Research Data | University System of New Hampshire." USNH Online Policy Manual. Accessed November 17, 2015. <https://www.usnh.edu/policy/unh/viii-research-policies/c-unh-policy-ownership-management-and-sharing-research-data>.
- University of Wisconsin-Madison. "Policy on Data Stewardship, Access, and Retention." Accessed March 4, 2016. <https://kb.wisc.edu/gradsch/page.php?id=34404>.
- Van Noorden, Richard. "Gates Foundation Announces World's Strongest Policy on Open Access Research." *Nature News Blog*, November 21, 2014. <http://blogs.nature.com/news/2014/11/gates-foundation-announces-worlds-strongest-policy-on-open-access-research.html>.
- Vines, Timothy H., Arianne Y. K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renault, and Diana J. Rennison. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24, no. 1 (January 2014): 94–97. doi:10.1016/j.cub.2013.11.014.
- Ward, Pauline. "Sources of Dataset Peer Review." DataShare wiki. Last modified January 12, 2016. <https://www.wiki.ed.ac.uk/display/dataset/Sources+of+dataset+peer+review>.
- Wellcome Trust. "Policy on Data Management and Sharing." Accessed October 21, 2014. <http://www.wellcome.ac.uk/about-us/policy/policy-and-position-statements/wtx035043.htm>.



## CHAPTER 3\*

# Collaborative Research Data Curation Services A View from Canada

*Eugene Barsky, Larry Laliberté, Amber  
Leahey, and Leanne Trimble*

In Canada, as in many developed countries, requirements for data management are being established across a wide range of scholarly disciplines. Barriers to data management and sharing are being addressed through the recommendation and use of community standards such as research data management plans (DMPs). Canada's federal granting agencies—known as the “Tri-Agencies,” consisting of the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC)—recently released a draft statement on digital data management.<sup>1</sup> Through this statement, the Tri-Agencies actively encourage research institutions to provide their researchers with an environment that enables robust stewardship and curation practices and to deliver support for the management and deposit of research data in secure, curated, and accessible repositories.

There are several library-led collaborative initiatives currently underway that aim to develop interoperable and sustainable data curation services in Canada in anticipation of future government requirements for data management. These initiatives, in combination with existing local expertise, are directly contributing to the capacity for research data management in Canadian universities. This chapter

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).



provides a brief history and overview of initiatives related to the coordination of data curation and preservation services at university libraries in Canada. Case studies from the Ontario Council of University Libraries (OCUL), the University of British Columbia Library (UBC), and the University of Alberta Libraries (UAL) are presented, with a focus on the library as a central facilitator of data curation and preservation. Some considerations about the financial and consortial business models are discussed. Finally, these efforts are placed in the context of Canada's overarching infrastructure initiative, the Canadian Association of Research Libraries (CARL) "Portage" project, which aims to develop a robust, collaborative national infrastructure network for Canadian research data.

## Canadian Academic Library Involvement in Research Data Management

Canada, like the United States, lacks a centralized data-archiving service. National data archives, like national libraries, provide government-supported services and expert staff to ensure that information produced within a country is permanently preserved. To date, there have been several attempts to establish a national data archive, but none have been able to secure adequate support or the funding required for its establishment.<sup>2</sup> Centralization tends to be challenging in a country that has a relatively small and geographically dispersed population characterized by regionalism. Nevertheless, libraries have been strong advocates for improved access to data in Canada. For example, the Canadian Association of Public Data Users (CAPDU, <http://www.capdu.ca/>) is a library-based organization whose members advocate for improved access to data in Canada. The Canadian Association of Research Libraries (CARL, <http://www.carl-abrc.ca/about-carl>) also has advocacy as part of its mandate and is involved in research data management activities. The efforts of Canadian academic librarians have seen success in strengthening the data collections available to researchers for secondary use.

The Data Liberation Initiative (DLI, <http://www.statcan.gc.ca/eng/dli/dli>), a subscription-based service providing access to Statistics Canada data, is an excellent early example of Canadian academic libraries collaborating on data management. The DLI program began in 1996 as a result of consultations between Statistics Canada, the Canadian Association of Research Libraries (CARL), and the Humanities and Social Sciences Federation of Canada.<sup>3</sup> The founding of the DLI was a response to both the high costs of Statistics Canada's public microdata files (which, due to budget cuts in the 1980s, were priced on a full cost-recovery

basis and were out of reach of all the most well-funded researchers) and the lack of data infrastructure at Canadian universities to provide access to these data.<sup>4</sup> The sheer size of the DLI collection, including thousands of data files for hundreds of survey series, and the demands from researchers for this data, has directly contributed to the growth of library data infrastructure to manage and preserve access to this data. When the DLI was formed, there was little expertise in many libraries to support data services; however, because Statistics Canada required a point of contact within the library who would be responsible for distributing data to end users, libraries quickly developed staff expertise through DLI training activities.<sup>5</sup> In addition, the DLI program prompted consortial initiatives to expand the available technical infrastructure. For example, in Ontario the development of <odesi> (<http://odesi.ca>) provided a centralized storage infrastructure and an innovative Web-based data access platform.

Some disciplines, particularly in the sciences, have developed a culture of data sharing through disciplinary repositories. In Canada, examples of domain repositories include the Polar Data Catalogue (a project of the Canadian Cryospheric Information Network, CCIN), the Canadian Astronomy Data Centre (an initiative of the Canadian Advanced Network for Astronomical Research, CANFAR), and CBRAIN (an initiative of the McGill Centre for Integrative Neuroscience, MCIN). Many disciplines, however, do not have these kinds of coordinated resources to turn to. Therefore a natural role for academic libraries is to develop institution-based data repositories and catalogues for disseminating and archiving data, particularly data sets that fall within the “long tail” of research data, meaning the large number of relatively small datasets that are produced in a wide range of disciplines.<sup>6</sup> Long-tail data sets have a great deal of diversity and can have high curation requirements. Libraries, with their expertise in preservation of research output (e.g., through institutional repositories) as well as their history of engagement in data management and dissemination activities, are well-equipped to take on these challenges, given sufficient resources.

The federal government has been consulting with various research communities, including libraries and archives, about the benefits and challenges of research data management for some time. In 2005, the Canadian government released the report of the *National Consultation on Access to Scientific Research Data* (NCASR), the cumulative work of an expert task force of more than seventy leaders Canada-wide from research, administration, and libraries, among other areas.<sup>7</sup> The list of recommendations included the development of a national steering body to coordinate data management and project funding across sectors in Canada; however, the approach ultimately failed to gain support politically.<sup>8</sup> In 2008, a new group was formed, the Research Data Strategy Working Group (RDSWG), that sought ways to move forward on the NCASR recommendations. In 2011, CARL and the RDSWG held a Research Data Summit, which resulted in the formation of Research Data Canada (RDC) in 2012.<sup>9</sup> RDC has facilitated a range of

committees and technical projects and partnered with other organizations internationally to advance research data infrastructure and expertise.

CARL has been an active participant in many of these important national discussions. In an effort to improve library preparedness for research data support services, it ran an extremely popular research data management course for libraries in early 2013.\* Building on the momentum generated by the course, a forum was established for ongoing dialogue around related activities in Canada, known as the Canadian Community of Practice for Research Data Management (RDM) in Libraries (<https://cancoprmd.wordpress.com/>). CARL has recognized that one of the ways forward for the library community is to establish more formal relationships with those organizations that provide Canada's research computing infrastructure, such as CANARIE (network infrastructure), Compute Canada (high performance computing), CUCCIO (chief information officers at Canada's universities), and the National Science Library (formerly known as CISTI, and the home of DataCite Canada).

Today academic libraries across Canada are putting plans in place to actively deliver a range of research data management services to their communities.<sup>10</sup> Infrastructure remains a central challenge, but one that is being addressed through collaborations between libraries and with the broader research community, through current CARL initiatives such as Portage. The Portage initiative brings together many stakeholders in a collaborative effort to develop distributed infrastructure, in contrast to earlier unsuccessful attempts to create a single national institution to manage data preservation. This bottom-up approach may be the key to success in the Canadian context.

## Overview of Case Studies

The authors of this chapter work at institutions across Canada that each has a unique approach to offering research data management services. Canada's small and spatially distributed population makes effective organization on a national level challenging. Canadian academic libraries tend to work together primarily within the context of regional consortia. In this chapter we will use several examples to illustrate the Canadian context. This chapter is not intended as a comprehensive description of all of the important research data management services undertaken at Canadian libraries, yet the case studies presented in this paper will show a good cross section of the kinds of research data management activities underway, ranging from libraries independently providing local services to comprehensive regional and national collaborations.

---

\* The outline from this course is available online as Canadian Association of Research Libraries, "Data Management Workshop," accessed August 3, 2016, <http://www.carl-abrc.ca/strengthening-capacity/workshops-and-training/data-management-workshop/>.

## *Local Services: University of Alberta Libraries*

The University of Alberta Libraries (UAL) has a long history of providing data services. In 1977, the precursor to the Data Library was established by data librarian Chuck Humphrey in University Computing Services (UCS), which ran a facility for data deposit and retrieval. The early Data Library started as a database registry of data sets generated by university researchers. By 1980 the database had grown into a full data library comprising local research data, such as the Edmonton Area Survey, as well as data obtained, through mediated access, from large data archives such as the ICPSR and The Roper Center. In 1992 the Data Library and its staff, a coordinator and a data librarian, became part of the libraries' Humanities and Social Sciences unit. The Data Library staff provided a full complement of research data support services, including data acquisition and cataloging, assistance with data analysis, instruction related to data, and the provision of data archiving services to university researchers. With the formation of the library's Digital Initiatives (DI) unit in 2012, the Data Library and its staff became part of a larger unit with a renewed focus on the development of new RDM services (<http://guides.library.ualberta.ca/data>).

Since 2014, a working group for Research Data Management Services (RDMS) has been coordinating services for the broader University of Alberta Libraries. The RDMS working group consists of ten members from various campus subject libraries, including health, sciences, and the humanities. The mandate of the working group is to develop an effective communication and outreach strategy for liaison librarians around research data management. To facilitate this role, the working group consults with librarians in order to provide them with the resources they need to provide information to their faculty in areas related to research data management. These resources include the collection of RDMS user stories reflecting these services and the development of a librarians' tool kit, which includes links to informational and educational resources and slide templates that can be modified and tailored to various teaching settings and levels.

One of the most prominent promotions of library services and training opportunities for researchers on the University of Alberta campus is the annual Research Data Management Week, which debuted as the Campus Data Summit in 2012. The week, also coordinated by the RDMS working group, is comprised of a mixture of keynotes, presentations, and workshops. The event is well attended, with over 200 attendees in 2015, and continues to thrive. In 2015, Compute Canada (<https://www.computecanada.ca/about/>) became heavily involved by offering a concurrent stream of workshops in order to introduce faculty to

Compute Canada's advanced research computing (ARC) systems, storage, and software, which provide services and infrastructure for Canadian researchers and their collaborators. The week also offers an opportunity to roll out new library services to a wide audience.

In 2014, the University of Alberta Libraries launched a Dataverse instance (<https://dataverse.library.ualberta.ca/dvn/>) to serve as an optional research data repository for the campus. Since the launch there have been many Dataverse workshops and one-off sessions for faculty and students; promotional slides and quick reference material have been added to the liaison librarian tool kit. As of March 2016, the UAL Dataverse contains thirty-four published Dataverses with 234 studies, 2,541 files and 1,986 downloads. There are also 115 unpublished Dataverses (many of which are ongoing projects).

Since 2015, the library has sponsored a data purchase program, noting that while open data is becoming more widely available, there are still many cases where data is available only commercially. Therefore, the libraries piloted a demand-driven data purchase program with the primary goal of purchasing data to better support University of Alberta researchers. Once the data is purchased, it is immediately made available to the researcher, and when the project is completed the data is added to Dataverse, provided the licensing allows for open distribution, for use by other interested campus researchers. If the licensing is restrictive, the files are still added to Dataverse for discoverability; however, access is mediated.

Finally, the Education and Research Archive (ERA), the University of Alberta's institutional repository, was developed and supported by the University of Alberta Libraries. ERA's open-access content includes the intellectual output of the university. In October 2015, all of ERA's content was migrated to a new Hydra-based digital asset management system (DAMS) environment. The new platform, called HydraNorth, is the first phase for consolidating all the diverse digital assets managed by the library. It currently harvests metadata from the Dataverse instance so that data sets can be discovered when users search ERA; then users are linked back to the data files in Dataverse via their persistent DOIs.

The UAL is on the leading edge of research data management services in Canadian academic libraries and serves as an excellent example of what can be achieved at universities with reasonable staffing and infrastructure funding. However, many Canadian universities may not have the resources to undertake these activities alone, and one solution is to seek opportunities to collaborate.

## *Informal Regional Consortia: University of British Columbia Library*

The University of British Columbia (UBC) Library is one of the largest university libraries in Canada and has been conducting ad-hoc research data management activities since the early 1970s. UBC Library's Abacus data repository (<http://dvn.library.ubc.ca/dvn/>) has, over the last fifteen years, moved from tape to custom database to a more complex data management system. In 2008, DSpace (version 1.5) was installed to run Abacus and replaced a home-grown system based on PHP and MySQL. As its input format was metadata-agnostic (using the Dublin Core metadata standard), it was suitable for the migration of UBC's licensed data sets, and the metadata management was the best available at the time of its adoption. Over time, the data needs of faculty and students increased dramatically. Data sets became larger and more complex. For example, geospatial data has gained wide use among research fields not normally associated with the use of data or geospatial imagery. The open-source software DSpace does not provide automatic version control, embedded data integrity checks, or granular access to data and data analysis in a web browser. As a result, the decision was made to upgrade UBC Abacus to a more data-user-friendly system, another open-source data repository solution, Dataverse.

Willing to assist smaller regional schools, in 2008, UBC entered into an arrangement to make the Abacus data repository available to other universities in the province. At the time of writing, four major university research libraries in British Columbia (Simon Fraser University, University of Victoria, University of Northern British Columbia, and University of British Columbia) are using the UBC instance of Dataverse, primarily as a licensed data repository. Using EZ-proxy for access control, data is provided to users from each institution according to their data licenses. Moreover, the UBC Abacus Dataverse has expanded to allow researchers from the universities to submit their open research data. Currently, UBC Abacus has more than 30,000 managed data files, with more than 10TB of managed data. The researcher-submitted data collection is approximately 10 percent of all data files but is steadily growing.

A UBC Library research data team provides basic and advanced Dataverse training to groups, departments, and labs on UBC campus as well as its partners in other university libraries and research institutes. After training, the goal is for these groups to manage their own data within the appropriate Dataverses assigned to them. The UBC team assumes responsibility for the entire Dataverse instance; however, individual researchers, labs, and libraries are trained and assigned to be the data curators for their own data sets.

## *Formal Regional Consortia: The Ontario Council of University Libraries*

In Ontario, several universities have a long history of providing data archiving services.\* The Carleton University Social Science Data Archive began in 1965 and was housed in the Sociology and Anthropology Department until around 1994, when it moved to the MacOdrum Library and become known as the Data Centre (now Data Services, <https://library.carleton.ca/contact/service-points/data-services>). The University of Western Ontario (now Western University) launched its Data Resources Library in the late 1970s (now known as the Map and Data Centre, <https://www.lib.uwo.ca/madgic/>), which worked with the Social Science Computing Laboratory to disseminate and archive several faculty research projects. The University of Toronto established its Data Library in 1988 (now the Map and Data Library, <http://mdl.library.utoronto.ca/>), with services that included the acquisition and preservation of data sets produced by University of Toronto researchers. By the late 1990s, as was happening across Canada after the initiation of the DLI program, additional universities in Ontario began to develop data expertise and to offer data support services to their communities.<sup>11</sup>

In Ontario, there are twenty-one universities, which vary widely in size, focus, and available resources. Since the 1960s, the libraries at these twenty-one universities had been collaborating through the Ontario Council of University Libraries (OCUL). In its early years, OCUL was involved in traditional library services such as consortial licensing of journals and facilitating effective resource sharing. In 2002 OCUL formed Scholars Portal (<http://www.scholarsportal.info/>), a shared technology infrastructure that hosts and provides access to OCUL's growing digital collections. As data services came to greater prominence, Ontario libraries saw an opportunity to collaborate under the OCUL umbrella in order to improve services, reduce duplication of effort, and better manage limited resources. Therefore, over the last decade, OCUL has undertaken several successful data infrastructure projects, including the development of <odesi>, a social science data portal, and Scholars GeoPortal (<http://geo.scholarsportal.info>), a geospatial data portal. While each of these does contain some research data, <odesi> and Scholars GeoPortal are intended as curated collections of “published” data sets from authoritative sources such as government statistical agencies and as such are not conducive to the widespread inclusion of member libraries’ institutional research data outputs. These systems are also primarily focused on discovery and access rather than long-term preservation.<sup>12</sup>

---

\* Canadian university data services are listed in this chronology (a work in progress) developed by members of the International Association for Social Science Information Services and Technology (IASSIST): “Chronology of Data Libraries and Data Centres,” accessed August 3, 2016, [https://docs.google.com/spreadsheets/d/1qmC\\_z50UDHh-3Jwldu6wGrtz1xjdfumBaad6P6cUpBdY/edit#gid=0](https://docs.google.com/spreadsheets/d/1qmC_z50UDHh-3Jwldu6wGrtz1xjdfumBaad6P6cUpBdY/edit#gid=0).



For this reason, other solutions were needed in Canada to address the growing demand for library research data repositories, and in 2011, Scholars Portal installed an instance of the Dataverse open-source software and offered it to the OCUL community as a pilot program. The pilot was intended to address a community-identified need for an Ontario-based repository service that would allow for easy-to-use, Web-based self-deposit by researchers. Dataverse was chosen for the pilot due to its support for research data, including the Data Documentation Initiative (DDI) metadata built in. Scholars Portal staff developed some documentation and training materials to inform and train staff at OCUL libraries about the benefits of incorporating Dataverse into the suite of services offered for data management and deposit of research data. As a result, the Scholars Portal Dataverse instance has allowed some OCUL libraries to launch research data management services without needing to have the technical infrastructure and staffing to support repositories of their own. Models for the service vary from library to library, ranging from self-serve deposit to library-mediated curation. Some examples of OCUL institutions that have launched research data management services based upon the Dataverse platform are the University of Guelph and Queen's University.<sup>13</sup> Due to the uptake of Dataverse within OCUL, the successful pilot became a core Scholars Portal service in 2012. Today, support for the use of Dataverse is largely provided by local library staff and is independent of the infrastructure hosted and supported by Scholars Portal.

In Ontario, several libraries have been offering longstanding RDM services, while others have recently embarked upon new RDM initiatives or are still in the planning phases. There is no doubt that this is a strategic area for most academic libraries, but it is unclear how RDM services will be funded at a time when budgets are very tight and researcher demand is in its infancy (with Canadian funder requirements still in flux). A community of librarians interested in research data management has begun to emerge, with the creation of an OCUL-wide Listserv to discuss topics of interest and an RDM theme for the 2015 Scholars Portal Day (<http://www.ocul.on.ca/node/4479>). Continued collaboration through the OCUL consortium will likely be extremely important to the success of emerging RDM services in Ontario libraries.

## Data Repository Services in Canadian Libraries

There are many factors that libraries must consider when selecting software to form the basis for a research data repository. A suite of software is needed that can support access and discovery as well as long-term preservation. Access and discovery are facilitated through support for established metadata standards and



harvesting protocols, granular search tools, and data exploration tools. Data preservation involves the ability to manage data identification (through persistent identifiers), integrity, sustainability, and authenticity.

## *Discovery and Access Platforms*

As we saw in the previous section, Dataverse has been the data repository software of choice for all of our example institutions. Dataverse (<http://dataverse.org>), developed by Harvard's Institute for Quantitative Social Science, is open-source software that allows researchers to share, cite, preserve, discover, and analyze research data.<sup>14</sup> Its open-source nature means that an institution or group of institutions can host its own instance of the Dataverse software and offer a customized solution tailored to its own community. This is an important factor in Canada, where many universities prefer to store data on local servers hosted within the country. A local installation also provides the opportunity for local branding and for offering custom training resources to users.

Dataverse is designed as a self-deposit platform, organized into Dataverse networks, where individual researchers, research teams, and institutes can create their own account and deposit their own data into "Dataverses" that are part of a bigger "network." It is also possible for university libraries or other data custodians to curate contributions and manage the data submission process on behalf of researchers. In this sense, Dataverse is very flexible. For example, in the University of Alberta Libraries' Dataverse, the entire network is devoted to research data from one institution, and an individual Dataverse is created for each research project being deposited. In British Columbia, the Abacus Dataverse Network focuses on library-curated Dataverses for each participating institution. In Ontario, the Scholars Portal Dataverse (<https://dataverse.scholarsportal.info/>) is completely open-ended, with some institutions hosting a library-curated Dataverse within the network, in addition to researcher-created Dataverses. Local branding is possible for both the network and each individual Dataverse contained within it.

Dataverse also provides data analysis functionality in the browser; users do not necessarily need to download the data files to interact with them. Tabular data files that are uploaded to the system can be further analyzed in the integrated web-based data analysis and visualization tool. Offering some data visualization and analysis within the Dataverse tool eliminates the need for desktop software to perform similar tasks and adds to the interactiveness of the data, potentially broadening the audience and range of users. Moreover, the Universal Numeric Fingerprint (UNF) feature in Dataverse works to enhance the reproducibility of science. A UNF "is a unique signature of the semantic content of a digital object. It is not simply a checksum of a binary data file. Instead, the UNF algorithm approximates and normalizes the data stored within. A cryptographic hash of that

normalized (or canonicalized) representation is then computed.<sup>15</sup> This means that same data object stored in, say, SPSS and Stata, will have the same UNF. And if the same analysis was used on the same data set, the UNF should be the same. Moreover, specific analyses done in Dataverse are given a special citation that mentions the analysis performed.

Dataverse is also easy to integrate with other library resources for improved discovery. For instance, since all partners with UBC Abacus Dataverse are using ProQuest's Summon as a discovery search engine for their libraries, the corresponding Dataverses are exposed via OAI protocol to their Summon engines. Each OAI feed includes all research data for the partner institutions and appropriate licensed data for that institution.\* Improved discovery (especially when assigning DOIs for research data sets) means that curated data could be easily accessed and reused by researchers (e.g., in ORCID, Google, Datacite, VIVO, Crossref, and other services), thereby enhancing citations and improving research metrics for individuals and institutions.

Dataverse has proven to be a flexible platform that can support many models for library RDM services. It offers a range of features that may improve data discoverability and access. It also does a good job of managing data files from a preservation perspective, such as managing versions, conducting checksums to maintain data integrity, and supporting persistent identifiers, such as handles and DOIs. Dataverse is capable of normalizing tabular data files into an ASCII text format with a companion DDI metadata record, which is considered a best practice for long-term preservation.<sup>16</sup> However, Dataverse is not a fully featured digital preservation system. It is format-agnostic and will accept deposit of all file types (not just tabular data), but currently it does not support normalization or metadata extraction from nontabular data files. The library community is in need of a robust long-term preservation solution that can manage a larger range of file formats and establish normalization and migration best practices for them. This preservation system would be used in conjunction with the established Dataverse service.

## *Long-Term Preservation*

Digital preservation activities are designed to secure the long-term future of digital information resources. A successful digital preservation strategy must account for and mitigate the impact of various threats to the accessibility and usability of digital materials over time. Common challenges include software, hardware, and media format obsolescence; hardware failure; and natural disasters, among many

---

\* An example of an OAI feed is available as University of British Columbia Libraries, Summon search result for "DBID: BAXLO," accessed August 3, 2016, <http://ubc.summon.serialssolutions.com/#!/search?ho=t&q=DBID:%20BAXLO&l=en>.

others. Mitigation strategies may include storage refresh, file format normalization (to open formats), software and hardware migration, data replication, and emulation.<sup>17</sup> Preservation metadata about the original data file, its provenance, and the preservation actions taken on the data (such as data validation or normalization to another file format) are required and therefore desired functionality for long-term preservation systems. Ensuring that that preservation activities are documented and well understood is crucial to ensuring long-term viability of data.

One software tool that has emerged in recent years to support digital preservation is Archivematica (<https://www.archivematica.org/en/>). Archivematica is an open-source software package developed by Artefactual Systems. It takes a “micro-services” approach to preservation, offering an integrated suite of free and open-source tools that allow users to process digital objects by applying format-specific preservation policies in order to prepare objects for archiving and dissemination.<sup>18</sup> Archivematica is essentially a pipeline of services that moves digital information packages through a series of file-system directories. Together these steps process digital objects from ingest to dissemination, resulting in the production of an Archival Information Package (AIP), a Dissemination Information Package (DIP), or both. An AIP is a container holding all the information necessary for long-term preservation of the file; it typically includes the original files and existing metadata, any normalized files created by Archivematica processes, and a preservation metadata file generated by Archivematica. This preservation metadata follows the PREMIS preservation metadata standard, encoded in METS (Metadata Encoding and Transmission Standard) format.\* In contrast, a DIP is a package delivered to an access platform and contains the data and metadata needed for discovery. Once created, AIPs and DIPs exist independently from Archivematica and are typically stored in a digital asset management system (DAMS) or other secure storage location. Used together, the Archivematica micro-services make it possible to fully implement the Open Archival Information System (OAIS) reference model, a framework for understanding the responsibilities and processes involved in the design of a preservation system.<sup>19</sup>

Digital preservation can be applied to all forms of digital information, including research data. Some work has been done to determine optimal file formats for statistical, geospatial, and other research data,<sup>20</sup> and Archivematica is equipped to handle relevant normalizations for a wide range of file formats, including images, spreadsheets, documents, and many other files. Archivematica maintains a Format Policy Registry (based on formats documented in the PRO-

---

\* For more information on PREMIS, see Priscilla Caplan, *Understanding PREMIS* (Washington, DC: Library of Congress, 2009), <http://www.loc.gov/standards/premis/understanding-premis.pdf>. The Library of Congress also provides information on using PREMIS with METS: “Using PREMIS with METS,” Library of Congress, October 15, 2010, <http://www.loc.gov/standards/premis/premis-mets.html>.

NOM format registry), which documents the actions the software can apply to specific file formats.\* For example, JPGs are identified as “jpeg image format” and are normalized to TIFF. Archivemata will store the original JPG and the derived TIFF in the AIP, referencing the original and converted file names and locations, and will use the PREMIS vocabulary to describe this normalization in the METS file. There are still many specialized file formats for which normalization tools do not exist and that are not yet described in registries like PRONOM. However, as additional information is acquired and new tools developed, Archivemata is well equipped to integrate new policies. This is an area being explored by libraries within Canada (as part of the Portage project) and elsewhere.<sup>21</sup>

Our institutions have varying degrees of engagement with digital preservation using tools like Archivemata. University of British Columbia (UBC) has engaged Archivemata as its digital preservation system since 2014, hosting the software in UBC’s EduCloud cloud-computing service.<sup>22</sup> Importantly for British Columbia, this service meets provincial privacy requirements under the Freedom of Information and Protection of Privacy Act. In addition, EduCloud offers the benefits of a virtual server hosting service, such as server consolidation, resource pooling, high service availability, and regular backups. At this time, three (out of four) UBC Library digital repositories are connected to Archivemata for digital preservation: DSpace (UBC cIRcle), CONTENTdm, and AtoM.

OCUL also has significant experience with digital preservation, having received Trustworthy Digital Repository Certification (TRAC) for its electronic journal repository in 2013.<sup>23</sup> Like UBC, OCUL has also been developing a private cloud storage service, known as the Ontario Library Research Cloud (OLRC, <http://www.ocul.on.ca/node/2126>), being rolled out in late 2015. While not actively using Archivemata at this time, OCUL is undertaking several initiatives to add new functionality to Archivemata, in collaboration with Artefactual Systems, in order to assess the opportunity to incorporate it as a service for OCUL libraries. Scholars Portal’s in-house solution for preservation of electronic journal content is not designed for self-serve access by individual OCUL member institutions for the preservation of their own local content (e.g., digitized collections). Scholars Portal staffing is not sufficient to manage local preservation activities on behalf of member institutions, nor is this considered desirable. Instead, Scholars Portal sees the combination of Archivemata and OLRC as a potential self-serve Web-based solution for supporting local preservation requirements. To this end, Scholars Portal is involved with integrating Archivemata with OpenStack Swift storage, the technology upon which OLRC is based. In addition to storage integration, a number of libraries across Canada (including UBC, University

---

\* Archivemata’s Format Policy Registry is described at <https://www.archivemata.org/en/docs/fpr/>, and the PRONOM registry at <https://www.nationalarchives.gov.uk/PRONOM/Default.aspx>.

of Alberta, and OCUL) are currently engaged under the Portage umbrella in a project to integrate Dataverse and Archivematica. When completed, this project will provide new opportunities for integrating good preservation practices into library research data repository workflows.

## Operational Costs of Data Repository Services

The costs of operating of a data repository can vary widely depending on the level of services provided, but in all cases there will be technology (hardware, software, and storage) and staffing costs. The use of open-source software like Dataverse eliminates the cost of software licensing fees; however, it can become necessary to invest software development resources in order to implement desired features in the software, as we will describe with an example in the Future Directions section.

The University of Alberta has taken on the operational costs of running Dataverse locally. The service is directly supported by four staff members, in addition to their other duties, who not only manage the technical infrastructure but also provide data curation services to researchers, including one-to-one consultation sessions on metadata creation, file permissions, the value of data sharing, and the importance of data attribution. Most of the technical implementation work was up front to get the service out the door, and episodic during software updates. Once the service was up and running, any operational costs related to its promotion (presentations/workshops) have been spread out to all librarians with portfolios relating to RDMS.

When university libraries work together in consortia (as it is frequently done in collections management), it is possible to share costs and reduce duplication of effort. British Columbia's Abacus Dataverse Network is an example of a collaborative service that is still in its early days. Since the collaborative work led by the University of British Columbia Library does not function as a formal consortium, it has been challenging to formalize a cost-sharing model; such models are not common in the province of British Columbia's academic libraries to date. However, it is not sustainable for UBC Library to continue paying for both the technical and human side of the operation, which in 2015 ran around \$250,000 CAD.

In Ontario, where there is a long-standing history of cost sharing through formal consortia, the growing pains are fewer. OCUL has an established model where new services are proposed to the governing group composed of the library directors for each member university. If the proposal is feasible and fits within OCUL's strategic directions, then OCUL will typically seek grant funding to

cover any one-time project costs such as development of a new software platform. When the service nears its launch date, the OCUL directors review a sustainability plan and make a decision as to whether to include this new service in the suite of “core services.” Once a service is considered a core service, it is integrated into the OCUL costing model, which calculates the contribution each member institution makes towards the OCUL annual budget.

In the case of the Scholars Portal Dataverse service, the model has been somewhat less formal. Because there was no new software to develop, grant funding was not sought. Also, the service was initially launched as a pilot with Scholars Portal assuming the up-front hardware costs, which were minimal at that time as the service was being used primarily for testing. To date, Scholars Portal staff have taken on a primarily technical support role for its Dataverse instance; users in need of more in-depth support for their data management activities are referred to designated staff at their home institution’s library. This differentiation of roles allows for technology-related costs to be centralized and shared among the OCUL consortium members, while research support costs are incurred by individual libraries as local expertise is needed. Today the OCUL Dataverse service is no longer considered a pilot, but the overall use of the service is still in its early growth phase. A sustainability plan is needed to establish requirements for data storage, staffing and resources for curation support services, and ongoing development projects, such as new features to meet local institutional or disciplinary needs. Additionally, OCUL has yet to finalize a costing model for long-term preservation of research data from member institutions.

## National Collaboration: Portage

In 2015, the Canadian Association of Research Libraries (CARL) launched the Portage network, an initiative to develop a library-based research data management network in Canada (<https://portagenetwork.ca>). The aim of Portage is to coordinate and expand existing expertise, services, and infrastructure so that all academic researchers in Canada will have access to the support they need for research data management. The goals of Portage are two-fold:

1. To develop and support national infrastructure platforms for planning, preserving, and discovering research data.
2. To provide services to researchers and related stakeholders through a national library-based network of expertise on research data management (RDM).<sup>24</sup>

Canada’s challenges in organizing nationally to support research data management and preservation has changed significantly in recent years. There is much greater awareness among funding agencies, campus research offices, and research-

ers themselves of the importance of data sharing and preservation. In addition, individual libraries have made inroads in supporting research data management locally and have positioned themselves as important partners in this area. The timing seemed right for something like Portage to bring about something that the library community has long desired: a national data archive.

## *Goal 1: Portage National Data Preservation Infrastructure*

The Portage initiative has participated in a series of pilot projects involving partners from within and beyond the library community through RDC's Federated Pilot initiative (<http://www.rdc-drc.ca/activities/federated-pilot/>). In particular, three projects have been central, and all of them have involved collaboration between Portage and Compute Canada. The goal has been to test a number of possible software stacks for ingesting data from a range of research data repositories (both institutional and disciplinary) into a distributed national preservation infrastructure. One project under the this umbrella,\* currently underway and described here, aims to integrate Dataverse and Archivematica, with the involvement of participants from across Canada, including OCUL's Scholars Portal, the University of British Columbia, the University of Alberta, Simon Fraser University, Artefactual Systems, and Dataverse.

The Dataverse-Archivematica integration has involved the development of customized open-source middleware that pulls published data sets from Dataverse instances using API calls and processes them for ingestion into Archivematica.<sup>25</sup> This involves the creation of a Submission Information Package (SIP), which combines a METS file describing the contents of the transfer, with the associated data files and metadata.<sup>26</sup> The middleware then initiates the ingest of the SIP into Archivematica. Processing the ingested content through the Archivematica pipeline is configured by the user on a case-by-case basis and therefore not part of the middleware. This middleware is under development for v4.x of Dataverse and is intended to be straightforward to update as Dataverse evolves

---

\* Another project under the Federated Pilot umbrella, spearheaded by Simon Fraser University and Compute Canada, integrated Islandora and Archivematica (Melissa Anez, "Archidora," DuraSpace wiki, last modified by Tim Hutchinson October 2, 2015, <https://wiki.duraspace.org/display/ISLANDORA715/Archidora>). A third explored integrating Archivematica with Globus Data Publication, a new tool that is already in use by Compute Canada. Some background information about all of these projects is available in a presentation given at the 2015 CNI Meeting (Martha Whitehead, Brian Owen, Dugan O'Neil, Leanne Trimble, and Geoff Harder, "Collaborating to Develop and Test Research Data Preservation Workflows" [slides from presentation, CNI Spring 2015 Membership Meeting, Seattle, WA, April 13–14, 2015], [https://www.cni.org/wp-content/uploads/2015/05/CNI\\_Collaborating\\_Whitehead.pdf](https://www.cni.org/wp-content/uploads/2015/05/CNI_Collaborating_Whitehead.pdf)).



(updates will not require any changes to the Archivematica software, only the middleware).

The overall goal of all of these related projects has been to generate a proof of concept that, through open standards and software, it is possible to ingest research data from a range of data repositories, perform preservation actions on the incoming data, and store the data in a distributed network that can accommodate a range of data types and storage locations. These initial pilot projects have shown promise, though scalability remains a concern. Portage is now working with Compute Canada on a set of requirements for a production platform, which would also integrate access and discovery as well as preservation. The focus for the next two years is on digital preservation and enhanced data discovery mechanisms, with an emphasis on building and improving open-source tools to enable curation and preservation of research data in Canada.

## *Goal 2: Portage Network of Expertise*

The Portage network of expertise is still in its infancy, but its operational goals and service model have been laid out in the network's organizational framework.<sup>27</sup> It is anticipated that the network will bring together expertise in key areas such as metadata, curation, access and dissemination, preservation, data management planning, security and confidentiality, and others. The first expert group formed was the Data Management Plan (DMP) Experts Group, tasked with developing the general data stewardship template to be included in a new Portage online tool, known as DMP Assistant (<https://assistant.portagenetwork.ca/>), for creating data management plans.

DMP Assistant is based upon the open-source DMPonline software created by the Digital Curation Centre in the United Kingdom (<https://dmponline.dcc.ac.uk/>) and is hosted at the University of Alberta. This tool is customized to meet Canadian needs with a bilingual interface and a standard DMP template developed in anticipation of the introduction of required data management plans by Canadian research councils. As funding agencies determine their requirements and research communities in Canada articulate the data planning needs that best fit their disciplinary profiles, templates will be incorporated within DMP Assistant to accommodate each new requirement.

In addition to developing the tool, the DMP Experts Group conducted usability tests with researchers and other stakeholders. As a result, the tool not only incorporates best practices in data stewardship, it also provides an easy-to-follow workflow that walks researchers through key questions about data management. Such plans typically identify how researchers will address data security, metadata production, file formats, file handling conventions, data sharing practices, data dissemination methods, and arrangements for long-term preservation.



## Future Directions

While the United States has seen data management planning requirements since 2011, which have been a strong driver for research data management activities,<sup>28</sup> Canadian efforts have been more anticipatory rather than reactive. For this reason it has been challenging at times to move forward with infrastructure development. Regardless, significant strides have been made and collaborations have been key to success in Canada to date. Many Canadian institutions are involved in RDM infrastructure projects at the local, provincial, or national level. There is a sense of momentum in this area, which must continue to build. But there is much more still to be done.

For example, in order for RDM infrastructure to meet the needs of all Canadian researchers, our user interfaces must be bilingual, since both English and French are official languages in Canada. The Portage DMP tool is an excellent example of new infrastructure being designed with this in mind. However, our data repository tools must follow. A project is underway to accomplish this for the Harvard-based open-source Dataverse software, where Scholars Portal staff are code contributors and are working on internationalizing the code (a project of interest to a number of other countries around the world as well). For example, the Université de Montréal in Québec has undertaken translation of the user interface text from English into French. Once this work is complete, this code may become part of the public Dataverse codebase and available to Dataverse instances around the world.

We anticipate that many projects of this nature will be undertaken under the umbrella of the Portage network. Together, it is hoped, these will come together to form the needed infrastructure for managing and preserving research data on a national level.

## Conclusions

It is an exciting time in Canada for research data management. Libraries are seeing new opportunities to engage with their communities and with one another. Along with these new opportunities inevitably come challenges, such as costly digital infrastructure that must be managed on an ongoing basis. A number of approaches to research data management infrastructure have been explored in Canada to date, but no one approach holds all the answers. The Portage project has great potential to meet some significant unmet needs but will need sustainable funding in order to be successful.

The development of open-source tools, infrastructure, and support services for research data management is crucial if Canadian scholars are to successfully integrate these new activities into their workflows. While formal funder require-

ments for data management planning or data sharing are not yet established in Canada, consultations are underway and requirements are expected. Academic libraries have a history of supporting data access, dissemination, and preservation as well as an established mandate to participate in the preservation of the research outputs of their community (e.g., in institutional repositories).<sup>\*</sup> Libraries can provide leadership around the adoption of best practices and open standards and partner with a range of stakeholders in the development of infrastructure and tools. In Canada, the library community has been extremely active in encouraging research data sharing, going back as far as the 1960s, and is well positioned to play a leadership role going forward.

## Notes

1. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada, "Draft Tri-Agency Statement of Principles on Digital Data Management," July 9, 2015, last modified July 20, 2015, <http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1>.
2. Chuck Humphrey, "Canada's Long Tale of Data," *Preserving Research Data in Canada* (blog), December 5, 2012, <http://preservingresearchdataincanada.net/2012/12/05/hello-world/>.
3. Chuck Humphrey and Elizabeth Hamilton, "Is It Working? Assessing the Value of the Canadian Data Liberation Initiative," *Bottom Line* 17, no. 4 (2004): 138, doi:10.1108/08880450410567428.
4. Ernie Boyko and Wendy Watkins, *The Canadian Data Liberation Initiative: An Idea Worth Considering?* IHSN Working Paper No 006, International Household Survey Network, November 2011, 2, <http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP006.pdf>.
5. Chuck Humphrey, "Collaborative Training in Statistical and Data Library Services," *Resource Sharing Information Networks* 18, no. 1–2 (2005): 167–81, doi:10.1300/J121v18n01\_13.
6. P. Bryan Heidorn, "Shedding Light on the Dark Data in the Long Tail of Science," *Library Trends* 57, no. 2 (2008): 280–99, [http://muse.jhu.edu/journals/library\\_trends/v057/57.2.heidorn.html](http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html).
7. David F. Strong and Peter B. Leach, *National Consultation on Access to Scientific Research Data: Final Report* (Canada: Task Force for the National Consultation on Access to Scientific Research Data, 2005), [https://www.cni.org/wp-content/uploads/2013/03/CNI\\_National\\_Newton.pdf](https://www.cni.org/wp-content/uploads/2013/03/CNI_National_Newton.pdf).
8. Humphrey, "Canada's Long Tale of Data."
9. Chuck Humphrey, "Community Actions to Preserve Research Data in Canada," *Preserving Research Data in Canada* (blog), December 11, 2012, <https://preservingresearch-dataincanada.net/2012/12/11/community-actions-to-preserve-research-data-in-canada/>.

---

<sup>\*</sup> See, for instance, "UBC Library Strategic Plan 2010–2015," accessed May 20, 2016, [https://about.library.ubc.ca/files/2012/09/StrategicPlan\\_2010.pdf](https://about.library.ubc.ca/files/2012/09/StrategicPlan_2010.pdf). Many similar examples exist.

10. Michael Steeleworthy, "Research Data Management and the Canadian Academic Library: An Organizational Consideration of Data Management and Data Stewardship," *Partnership* 9, no. 1 (2014): 1–11, <https://journal.lib.uoguelph.ca/index.php/perj/article/view/2990/3278>.
11. Humphrey, "Collaborative Training in Statistical and Data Library Services."
12. Erin Forward, Amber Leahey, and Leanne Trimble, "Shared Geospatial Metadata Repository for Ontario University Libraries: Collaborative Approaches," *New Review of Academic Librarianship* 21, no. 2 (2015): 170–84, doi:10.1080/13614533.2015.1022662.
13. Wayne Johnston, "Digital Preservation Initiatives in Ontario: Trusted Digital Repositories and Research Data Repositories," *Partnership* 7, no. 2 (2012): 1–8, <http://criticalvoices.lib.uoguelph.ca/index.php/perj/article/view/2014/2637>; Jeff Moon, "Developing a Research Data Management Service: A Case Study," *Partnership* 9, no. 1 (2014): 1–14, <http://criticalvoices.lib.uoguelph.ca/index.php/perj/article/view/2988/3266>.
14. Mercè Crosas, "The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data," *D-Lib Magazine* 17, no. 1/2 (2011), doi:10.1045/january2011-crosas.
15. Dataverse, "Universal Numerical Fingerprint (UNF)," accessed March 29, 2016. <http://guides.dataverse.org/en/4.2.4/developers/unf/index.html>.
16. Claire Austin, Susan Brown, Chuck Humphrey, Amber Leahey, and Peter Webster, *Guidelines for the Deposit and Preservation of Research Data in Canada* (Ottawa: Research Data Canada, 2015), <http://www.rdc-drc.ca/wp-content/uploads/Guidelines-for-Deposit-of-Research-Data-in-Canada-2015.pdf>.
17. Brian Lavoie and Lorcan Dempsey, "Thirteen Ways of Looking at... Digital Preservation," *D-Lib Magazine* 10, no. 7/8 (July/August 2004), <http://mirror.dlib.org/dlib/july04/lavoie/07lavoie.html>.
18. Peter Van Garderen, "Archivematica: Using Micro-Services and Open-Source Software to Deliver a Comprehensive Digital Curation Solution," *Proceedings of the 7th International Conference on Preservation of Digital Objects (iPRES2010)*, 145–49, <https://ipres-conference.org/ipres10/papers/vanGarderen28.pdf>.
19. Ibid.
20. Guy McGarva, Steve Morris, and Greg Janée, *Technology Watch Report: Preserving Geospatial Data*, DPC Technology Watch Series Report 09-01 (Digital Preservation Coalition, May 2009), [http://www.dpconline.org/component/docman/doc\\_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee](http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee); Inter-university Consortium for Political and Social Research, *Principles and Good Practices for Preserving Data*, IHSN Working Paper No 003, International Household Survey Network, December 2009, <http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP003.pdf>; Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections," accessed November 17, 2015, <http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml>.
21. Jenny Mitcham, Chris Awre, Julie Allinson, Richard Green, and Simon Wilson, "Filling the Digital Preservation Gap: A Jisc Research Data Spring Project: Phase One Report—July 2015," accessed November 12, 2015, [http://figshare.com/articles/Filling\\_the\\_Digital\\_Preservation\\_Gap\\_A\\_Jisc\\_Research\\_Data\\_Spring\\_project\\_Phase\\_One\\_report\\_July\\_2015/1481170](http://figshare.com/articles/Filling_the_Digital_Preservation_Gap_A_Jisc_Research_Data_Spring_project_Phase_One_report_July_2015/1481170).
22. Bronwen Sprout and Mark Jordan, "Archivematica As a Service: COPPUL's Shared Digital Preservation Platform/Le service Archivematica: La plateforme partagée de

- conservation de documents numériques du COPPUL,” *Canadian Journal of Information and Library Science* 39, no 2 (2015): 235–44, <https://open.library.ubc.ca/cIRcle/collections/ubclibraryandarchives/494/items/1.0132717>.
23. Center for Research Libraries, *Report on Scholars Portal Audit* (Chicago: Center for Research Libraries, February 2013), <https://www.crl.edu/reports/scholars-portal-audit-report-2013>.
  24. Portage Network, “Governance Structure,” *Portage website*, accessed July 21, 2016, <https://portagenetwork.ca/about/governance-structure/>.
  25. Artefactual Systems, “Dataverse,” Archivemata Wiki, accessed November 12, 2015, <https://wiki.archivemata.org/Dataverse>.
  26. Ibid.
  27. Whitehead and Shearer, *Portage*, 2–3.
  28. Katherine G. Akers, Fe C. Sferdea, Natsuko H. Nicholls, and Jennifer A. Green, “Building Support for Research Data Management: Biographies of Eight Research Universities,” *International Journal of Digital Curation* 9, no. 2 (2014): 171–91, doi:10.2218/ijdc.v9i2.327.

## Bibliography

- Akers, Katherine G., Fe C. Sferdea, Natsuko H. Nicholls, and Jennifer A. Green. “Building Support for Research Data Management: Biographies of Eight Research Universities.” *International Journal of Digital Curation* 9, no. 2 (2014): 171–91. doi:10.2218/ijdc.v9i2.327.
- Anez, Melissa. “Archidora.” *DuraSpace wiki*. Last modified by Tim Hutchinson October 2, 2015. <https://wiki.duraspace.org/display/ISLANDORA715/Archidora>.
- Archivemata. “Format Policy Registry.” Accessed August 2, 2016. <https://www.archivemata.org/en/docs/fpr/>.
- Artefactual Systems. “Dataverse.” Archivemata Wiki. Accessed November 12, 2015. <https://wiki.archivemata.org/Dataverse>.
- Austin, Claire, Susan Brown, Chuck Humphrey, Amber Leahey, and Peter Webster. *Guidelines for the Deposit and Preservation of Research Data in Canada*. Ottawa: Research Data Canada, 2015. <http://www.rdc-drc.ca/wp-content/uploads/Guidelines-for-Deposit-of-Research-Data-in-Canada-2015.pdf>.
- Boyko, Ernie and Wendy Watkins. *The Canadian Data Liberation Initiative: An Idea Worth Considering?* IHSN Working Paper No 006. International Household Survey Network, November 2011. <http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP006.pdf>.
- Canadian Association of Research Libraries. “Data Management Workshop.” Accessed August 3, 2016. <http://www.carl-abrc.ca/strengthening-capacity/workshops-and-training/data-management-workshop/>.
- Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada. “Draft Tri-Agency Statement of Principles on Digital Data Management.” July 9, 2015. Last modified July 20, 2015. <http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1>.

- Caplan, Priscilla. *Understanding PREMIS*. Washington, DC: Library of Congress, 2009. <http://www.loc.gov/standards/premis/understanding-premis.pdf>.
- Center for Research Libraries. *Report on Scholars Portal Audit*. Chicago: Center for Research Libraries, February 2013. <https://www.crl.edu/reports/scholars-portal-audit-report-2013>.
- Crosas, Mercè. "The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data." *D-Lib Magazine* 17, no. 1/2 (2011). doi:10.1045/january2011-crosas.
- Dataverse. "Universal Numerical Fingerprint (UNF)." Accessed March 29, 2016. <http://guides.dataverse.org/en/4.2.4/developers/unf/index.html>.
- Forward, Erin, Amber Leahey, and Leanne Trimble. "Shared Geospatial Metadata Repository for Ontario University Libraries: Collaborative Approaches." *New Review of Academic Librarianship* 21, no. 2 (2015): 170–84. doi:10.1080/13614533.2015.1022662.
- Harder, Geoff, Leanne Trimble, Dugan O'Neil, Brian Owen, and Martha Whitehead. "Collaborating to Develop and Test Research Data Preservation Workflows." Paper presented at the CNI Spring 2015 Meeting, Seattle, WA, April 13–14, 2015. <https://www.cni.org/topics/ci/collaborating-to-develop-and-test-research-data-preservation-workflows>.
- Heidorn, P. Bryan. "Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends* 57, no. 2 (2008): 280–99. [http://muse.jhu.edu/journals/library\\_trends/v057/57.2.heidorn.html](http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html).
- Humphrey, Chuck. "Canada's Long Tale of Data." *Preserving Research Data in Canada* (blog). December 5, 2012. <http://preservingresearchdataincanada.net/2012/12/05/hello-world/>.
- . "Collaborative Training in Statistical and Data Library Services." *Resource Sharing Information Networks* 18, no. 1–2 (2005): 167–81. doi:10.1300/J121v18n01\_13.
- . "Community Actions to Preserve Research Data in Canada" *Preserving Research Data in Canada* (blog). December 11, 2012. <https://preservingresearchdataincanada.net/2012/12/11/community-actions-to-preserve-research-data-in-canada/>.
- Humphrey, Chuck, and Elizabeth Hamilton. "Is It Working? Assessing the Value of the Canadian Data Liberation Initiative." *Bottom Line* 17, no. 4 (2004): 137–46. doi:10.1108/08880450410567428.
- International Association for Social Science Information Services and Technology (IASSIST). "Chronology of Data Libraries and Data Centres." Accessed August 3, 2016. [https://docs.google.com/spreadsheets/d/1qmC\\_z50UDHh3Jwdu6wGrtz1xjdfumBaad6P-6cUpBdY/edit#gid=0](https://docs.google.com/spreadsheets/d/1qmC_z50UDHh3Jwdu6wGrtz1xjdfumBaad6P-6cUpBdY/edit#gid=0).
- Inter-university Consortium for Political and Social Research. *Principles and Good Practices for Preserving Data*. IHSN Working Paper No 003. International Household Survey Network, December 2009. <http://www.ihsn.org/home/sites/default/files/resources/IHSN-WP003.pdf>.
- Johnston, Wayne. "Digital Preservation Initiatives in Ontario: Trusted Digital Repositories and Research Data Repositories." *Partnership* 7, no. 2 (2012): 1–8. <http://criticalvoices.lib.uoguelph.ca/index.php/perj/article/view/2014/2637>.
- Lavoie, Brian, and Lorcan Dempsey. "Thirteen Ways of Looking at... Digital Preservation." *D-Lib Magazine* 10, no. 7/8 (July/August 2004). <http://mirror.dlib.org/dlib/july04/lavoie/07lavoie.html>.

- Library of Congress. "Sustainability of Digital Formats: Planning for Library of Congress Collections." Accessed November 17, 2015. <http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml>.
- McGarva, Guy, Steve Morris and Greg Janée. *Technology Watch Report: Preserving Geospatial Data*. DPC Technology Watch Series Report 09-01. Digital Preservation Coalition, May 2009. [http://www.dpconline.org/component/docman/doc\\_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-greg-janee](http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-greg-janee).
- Mitcham, Jenny, Chris Awre, Julie Allinson, Richard Green, and Simon Wilson. "Filling the Digital Preservation Gap: A Jisc Research Data Spring Project: Phase One Report—July 2015." Accessed November 12, 2015. [http://figshare.com/articles/Filling\\_the\\_Digital\\_Preservation\\_Gap\\_A\\_Jisc\\_Research\\_Data\\_Spring\\_project\\_Phase\\_One\\_report\\_July\\_2015/1481170](http://figshare.com/articles/Filling_the_Digital_Preservation_Gap_A_Jisc_Research_Data_Spring_project_Phase_One_report_July_2015/1481170).
- Moon, Jeff. "Developing a Research Data Management Service: A Case Study." *Partnership* 9, no. 1 (2014): 1–14. <http://criticalvoices.lib.uoguelph.ca/index.php/perj/article/view/2988/3266>.
- Portage Network. "Governance Structure." Accessed July 21, 2016, <https://portagenetwork.ca/about/governance-structure/>.
- PRONOM homepage. Accessed August 3, 2016. <https://www.nationalarchives.gov.uk/PRONOM/Default.aspx>.
- Sprout, Bronwen, and Mark Jordan. "Archivematica As a Service: COPPUL's Shared Digital Preservation Platform/Le service Archivematica: La plateforme partagée de conservation de documents numériques du COPPUL." *Canadian Journal of Information and Library Science* 39, no. 2 (2015): 235–44. <https://open.library.ubc.ca/cIRcle/collections/ubclibraryandarchives/494/items/1.0132717>.
- Steeleworthy, Michael. "Research Data Management and the Canadian Academic Library: An Organizational Consideration of Data Management and Data Stewardship." *Partnership* 9, no. 1 (2014): 1–11. <https://journal.lib.uoguelph.ca/index.php/perj/article/view/2990/3278>.
- Strong, David F., and Peter B. Leach. *National Consultation on Access to Scientific Research Data: Final Report*. Canada: Task Force for the National Consultation on Access to Scientific Research Data, 2005. [https://www.cni.org/wp-content/uploads/2013/03/CNI\\_National\\_Newton.pdf](https://www.cni.org/wp-content/uploads/2013/03/CNI_National_Newton.pdf).
- University of British Columbia Libraries. "Summon search result for DBID: BAX-LO." Accessed August 3, 2016. <http://ubc.summon.serialssolutions.com/#!/search?ho=t&q=DBID:%20BAXLO&l=en>.
- . "UBC Library Strategic Plan 2010–2015." Accessed May 20, 2016. [https://about.library.ubc.ca/files/2012/09/StrategicPlan\\_2010.pdf](https://about.library.ubc.ca/files/2012/09/StrategicPlan_2010.pdf).
- Van Garderen, Peter. "Archivematica: Using Micro-Services and Open-Source Software to Deliver a Comprehensive Digital Curation Solution." *Proceedings of the 7th International Conference on Preservation of Digital Objects (iPRES 2010)*, 145–49. <https://ipres-conference.org/ipres10/papers/vanGarderen28.pdf>.
- Whitehead, Martha, Brian Owen, Dugan O'Neil, Leanne Trimble, and Geoff Harder. "Collaborating to Develop and Test Research Data Preservation Workflows." CNI Spring 2015 Membership Meeting, Seattle, WA. April 13–14, 2015. [https://www.cni.org/wp-content/uploads/2015/05/CNI\\_Collaborating\\_Whitehead.pdf](https://www.cni.org/wp-content/uploads/2015/05/CNI_Collaborating_Whitehead.pdf).







## CHAPTER 4\*

# Practices Do Not Make Perfect

## Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation

*Ixchel M. Faniel and Elizabeth Yakel*

### Introduction

An unprecedented amount of data sharing and reuse is now possible, but disciplinary practices and traditions can create challenges for researchers wanting to meaningfully reuse data other researchers created for different purposes. Until recently in many disciplines, data sharing among peers occurred informally, in response to colleagues' requests. Now given the ability to generate digital data, federal mandates for data management and sharing, and the motivation to pose interdisciplinary questions that address critical social and environmental problems, some data producers are expected to formally share data via deposit into a repository with limited guidance about what to share and how to share it. However, the financial, technological, and human resources required to prepare data

---

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

for sharing are limited or nonexistent. Yet even with these challenges, data sharing and reuse are growing. We contend additional growth can occur with repository staff's increased understanding of their designated communities of data reusers.

In this chapter, we draw from the results of the Dissemination Information Packages for Information Reuse (DIPIR) project. A multiyear investigation jointly funded by the Institute of Museum and Library Services, OCLC, and University of Michigan, DIPIR has investigated data sharing and reuse practices within three academic communities: quantitative social science (i.e., social science), archaeology, and zoology. Over the years we identified a number of interesting similarities and contrasts across the disciplines with regard to data sharing and reuse.<sup>1</sup> In this chapter we focus on three areas: (1) disciplinary practices and traditions surrounding data sharing and reuse within the three communities, (2) researchers' development of trust in the data they seek to reuse, and (3) sources of contextual information researchers rely on in addition to the repository. In the sections that follow, we describe our research methodology, discuss our findings, and conclude by describing the implications of this study for repository practice.

## Overview and Methodology for the DIPIR Project

The DIPIR project aimed to identify significant factors affecting data reuse and to consider the implications they have for repository practice. We focused on the social science, archaeological, and zoological research communities because of the differences in their data sharing and reuse practices and the different repository infrastructures each had in place for archiving and disseminating data for reuse.

To conduct this research project we collaborated with three key individuals at three disciplinary repositories: (1) Nancy McGovern representing the social sciences discipline at the Inter-university Consortium for Political and Social Research (ICPSR), (2) Eric Kansa representing the archaeology discipline at Open Context, and (3) William Fink representing the zoology discipline at the University of Michigan Museum of Zoology (UMMZ). Although we used the repositories to gain access to data reusers, our research collaborators helped facilitate access to a broader disciplinary network of users beyond the repositories.

Our data collection plan employed a mixed-methods approach in that we used multiple data collection techniques, both qualitative and quantitative, to address our research questions. By using mixed methods, we were able to triangulate data from the different methods to more fully answer our research questions and address the limitations of each individual method. We conducted semi-structured interviews with data reusers in each discipline. We then employed a secondary data collection technique especially suited for each discipline (table 4.1). Specifi-

cally, we implemented a survey of social scientists because of the large population of data reusers in that area. We observed zoologists working with specimens in a museum because their research practice involves interactions with both physical specimens and digital repositories. Finally, we analyzed server logs from Open Context because we were interested in understanding how archaeologists who are new to data reuse navigated and worked with digital data. Mixed-methods studies are good for addressing complex environments and issues and can lead to increased validity and reliability.<sup>2</sup> Data were collected in the following ways:

- We recruited staff and data reusers for interviews from our three collaborating organizations. Interviewees also were recruited through disciplinary conferences and snowball sampling techniques, so in all cases our sample consisted of data reusers beyond our partner institutions.
- We surveyed social scientists using the *ICPSR Bibliography of Data-Related Literature* as our sample.<sup>3</sup> Social scientists were surveyed using Qualtrics an online survey application.
- We collected server logs from the Open Context repository between August 2011 and December 2013.
- Finally, we observed zoologists interacting with physical specimens at UMMZ.

This staged approach to data collection aligns with Creswell’s “sequential explanatory strategy” in which each different data collection method builds on the previous method to address broader research questions.<sup>4</sup>

**TABLE 4.1**  
DIPIR Data Collection Methods and Final Participant Numbers by Discipline

	Archaeology	Zoology	Social Science
<b>Phase 1: Project Start-Up (2011)</b>			
Staff Interviews	4	10	10
<b>Phase 2: Collecting and Analyzing Reuser Data (2011–2013)</b>			
Interviews	22	27	43
Observations		13	
Survey			237
Server log entries	572,134		

In this chapter, we focus on findings from the interviews and observations. To analyze these data, we began by coding transcripts from the interviews and observations using NVivo, a qualitative data analysis software package. We created an initial codeset that was based on the data reuse literature and the interview pro-

toloc. During analysis, team members discussed the coding and added codes that emerged from the data. Coding began with two DIPIR team members working on the same transcript so we could test for interrater reliability (IRR). We used Scott's pi to calculate IRR. Our IRR scores were 0.73 for the archaeologists, 0.74 for the interviews with zoologists, 0.88 for the observations of zoologists, 0.77 for the expert social scientists, and 0.88 for the novice social scientists. When IRR was achieved, each person coded transcripts independently. Once the data were fully coded we went through several phases of analysis to delve more deeply into the findings related to each code as well as to identify relationships between codes.

## Disciplinary Traditions for Data Sharing and Reuse

Increased interest in sharing and reusing data has several common drivers regardless of discipline. Computing power and communication bandwidth have enabled data to be generated, shared, and analyzed more easily and cheaply.<sup>5</sup> In addition, federal regulations and mandates have effectively mobilized attention and support for public access to the data and other research outputs. Since the OMB circular A-110 in 1999, federal funding agencies have issued data sharing mandates, required data management plans, and begun to allow budget items related to data management, preparation, and sharing.<sup>6</sup>

In response to a 2013 White House Office of Science and Technology Policy memorandum, many federal agencies have developed policies to increase public access to federally funded research outputs.<sup>7</sup> Initiatives within the higher education and research communities, such as SHARE (SHARED Access Research Ecosystem), have been established to facilitate university compliance and to better meet stakeholders' research needs.<sup>8</sup> In addition, academic and research libraries have begun to develop services to support researchers' data management, sharing, and curation needs.

Accompanying these drivers are large-scale, interdisciplinary research studies in the sciences and humanities, where data reuse is vital. For example, in our DIPIR work we saw archaeologists—who once focused on a single site—reusing data from multiple sites in quantities larger than any one person could collect in a lifetime in order to examine regional social, economic, and cultural transitions between ancient civilizations.<sup>9</sup> Zoologists conducting biodiversity research were reusing data from repositories such as GenBank and the Global Biodiversity Information Facility (GBIF) to address questions about extinction or migration events, and social scientists were integrating government and academic research data to study household economic trends over time. In the following paragraphs, we discuss specific disciplinary practices and traditions as they relate to data sharing and reuse.

## *Social Scientists*

Social scientists have the benefit of over fifty years of data sharing and reuse through repositories at institutions such as the ICPSR, the Howard W. Odum Institute for Research in Social Science, and the Roper Center for Public Opinion Research. The repositories curate data that tends to be well-structured and homogeneous; their data includes survey data, public opinion polls, administrative data, and international political, economic, and social indicators. The US federal government is one of the largest producers of social science data, followed by academic researchers, private survey and marketing firms, and research organizations.<sup>10</sup>

Given the longevity of social science data repositories, best practices in digital preservation and archiving have emerged. The Reference Model for an Open Archival Information System (OAIS) has been used as a guide, and the Data Seal of Approval has been awarded to several institutions as a signal that their repositories are trustworthy preservation archives, including ICPSR, the Odum Institute, and the Roper Center. The repository infrastructure has created a sound base for social scientists to build a disciplinary tradition around data sharing and reuse, but not without challenges.<sup>11</sup>

The DIPIR study revealed that in the social science community, data collection can be complex and dynamic, particularly for large-scale, longitudinal studies, which may involve a variety of sampling procedures, the attrition of survey respondents, and changes to survey questions over time. Privacy concerns also arise when collecting some types of personally identifiable data. However, our interviews suggested the repositories were well staffed and developed practices to address these issues. For instance, in some cases, ICPSR staff recruited data from major studies before the team's data collection had begun, which allowed the articulation of curation goals and a negotiation of needs to occur at the beginning of the data life cycle. In other cases, repository staff had long-standing relationships with data producers at various survey organizations, state and local governments, and federal agencies to archive data, which enabled a common understanding of needs to develop over time. Moreover, the social scientists studied were dealing with a select few data formats, so repositories could easily convert data into mainstream software packages (i.e., SPSS, Stata, SAS, Excel) as well as preservation friendly formats (i.e., CSV).

Building on the knowledge and experience within the community over time, data deposit and documentation requirements were explicit and detailed, such as in the case of the ICPSR. Codebooks evolved as a standard way of describing data within the community, and the DDI standard developed in turn as a way to compile, present, and exchange data documentation.<sup>12</sup> Moreover, research shows social scientists' satisfaction with data reuse is positively related to high-quality data documentation.<sup>13</sup> Given a long-standing culture of data sharing, a mature

repository infrastructure, and well-established relationships with data producers, data sharing and reuse have become well-established within the social science community.

## *Archaeologists*

Archaeologists face internal and external pressures to change their data sharing and reuse practices and traditions. In addition to government mandates, data collection and dissemination practices and publication norms are changing for other reasons that are driven by cultural and political factors, in reaction to previous large-scale removal of cultural property from the country of origin. Data sharing and reuse practices in archaeology also are being adopted at different rates depending on the different sub-areas of the discipline.<sup>14</sup> Furthermore, the repository infrastructure to support data sharing and reuse is only recently emerging.

Legal and ethical mandates affect archaeology more than social science. For example, international and national legislation against the removal of cultural property means that archaeologists must document artifacts on-site and that they no longer have the luxury of shipping items home for further analysis and study.<sup>15</sup> In addition, professional organizations, such as the Society of American Archaeology, mandate that authors provide a “Data Availability Statement” detailing the “disposition and accessibility of the physical and digital data on which the research is based.”<sup>16</sup> A final push toward new models of data sharing and reuse comes from the publishers. Traditionally, archaeologists have published books with large appendices listing artifacts, measurements, drawings of sites, and so on. Our interviewees noted that many publishers are no longer willing to print these, so archaeologists must identify other means for distributing these data tables, site information, and analyses. All of these factors have converged to move archaeology researchers into the early stages of practicing data sharing and reuse.

The DIPIR study identified two aspects of archaeological practice that make data sharing and reuse difficult. First, the variety of data types used to document an archaeological site presents a challenge. Archaeologists essentially destroy the context of field sites during excavation; therefore data collection best practice requires documentation of the physical surroundings in exhaustive detail.<sup>17</sup> As a result, archaeologists create and rely on different types of data (e.g., photographs, field notes, measurements) in a variety of formats; these types can range from hand-drawn maps and figures to proprietary files that require special software (e.g., CAD drawings and GIS shape files).<sup>18</sup> In turn, data reusers may need to contextualize the physical artifacts to a variety of analog and digital data as part of the reuse process.

Second, the DIPIR study revealed that archaeology lacks common data recording practices. The archaeological community has not yet developed a shared

understanding about the documentation and contextual information needed for data reuse, and there are few agreed-upon standards for data collection within the community.<sup>19</sup> Moreover, the lack of standards hampers any interoperability of archaeological data across sites and sometimes over the course of a single excavation when it takes place over a long period of time.<sup>20</sup> This makes data reuse difficult. In one study, several archaeologists who analyzed the same dataset reached different conclusions; despite inadequate documentation, each archaeologist concluded that the data were trustworthy enough to conduct the types of analyses they wanted to accomplish.<sup>21</sup>

The absence of standard data repository infrastructure also hampers data reuse in archaeology.<sup>22</sup> This applies to both the existence of trusted and sustainable repositories as well as agreed-upon standards for curation. There are no metadata standards to encode or encapsulate the different types of data collected in the discipline of archaeology, and there are no agreed-upon vocabularies or ontologies to link related materials.<sup>23</sup> This is especially problematic since archaeological data are dispersed worldwide. For example, artifacts may be in museums or remain at the discovery site, while field notes, images, and other documentation remain with the archaeologist or in a different type of repository. While museums traditionally house the physical artifacts, the availability of repositories to deposit the digital data collected in the field is relatively new in archaeology. For example, Open Context and The Digital Archaeological Record (tDAR) are both less than ten years old. In our interviews, archaeologists described performing separate searches across many different sites to look for reusable data.

## Zoologists

Zoologists have built a strong data sharing and reuse infrastructure over centuries,<sup>24</sup> but computerization and advanced analytical techniques recently have transformed the nature of research. Zoology has gone from an observational science where taxonomic identifications were made on the basis of visual inspection to a field where DNA is used to categorize and verify species. In parallel with this transformation, new standards for sharing data, such as Darwin Core (an expansion of Dublin Core for biological taxa), were developed and repositories began to emerge for data sharing and reuse.

The DIPIR study found that the repository infrastructure for data sharing is strong in zoology. In the past, both amateur naturalists and professional zoologists deposited physical specimens in museums.<sup>25</sup> In recent decades this practice has been formalized by requiring those individuals who are collecting specimens to get legal permits or licenses, which in turn mandate the deposit of physical specimens to a museum.<sup>26</sup> At the same time, zoologists (the primary data collectors) are affiliated with museums that have collection managers on staff to assist



with the preparation, management, and curation of the physical specimens and their accompanying data and documentation. Interviewees noted that having dedicated curatorial staff to create documentation as a part of the research workflow has made it easier for zoologists to share and reuse data.

Accepted standards for zoological data are entrenched in the repository infrastructures that span the different formats of specimen data—from labels on a physical specimen, to a Darwin Core metadata representation, to a DNA sequence. Data from individual museums are aggregated to national and international repositories, such as VertNet and GBIF. Standardization of metadata, particularly Darwin Core, has enabled a rich array of interconnected repositories with different metadata representations of the same specimen at various levels of granularity. Aggregating zoological collections makes data discovery and access more efficient, but the levels of metadata also vary, so provenance information that traces the different representations back to the museum that holds the original physical specimen is important. While basic specimen data are standardized often using Darwin Core (e.g., what was collected, who collected it, where was it collected, when was it collected), curating the deeper context is more difficult and requires access to other sources, such as field books or specimen images (including x-rays). For data reuse studies that require information beyond the basics, the lack of context can complicate the reuse process.

## Data Reuse and Trust

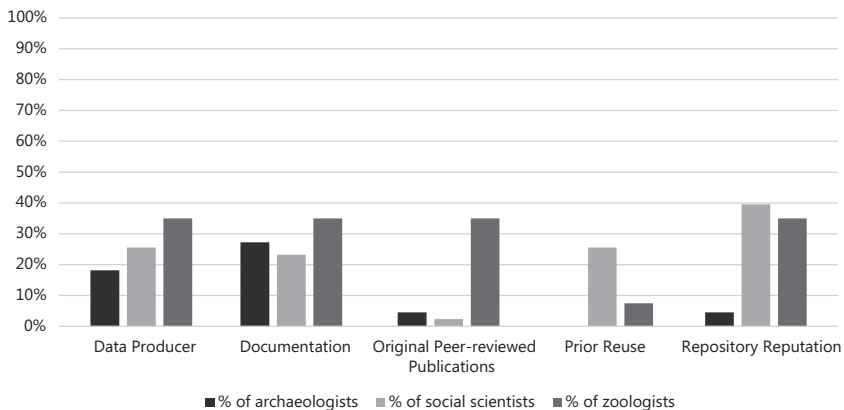
Trust in both the data and the repository plays a major role in whether or not data are reused. In the digital curation community, trust in repositories is often conceptualized in terms of the Trustworthy Repositories Audit & Certification (TRAC) process. TRAC is based on evaluating the internal processes of repositories and trust is synonymous “with ‘reliable’, ‘responsible’, ‘trustworthy’, and ‘authentic,’ in relation to archival functions such as creating, managing, and using digital objects.”<sup>27</sup> However, our DIPIR research was more interested in trust from data reusers’ points of view. Therefore, we adopted a more classic definition of trust as “a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another.”<sup>28</sup> We viewed trust as a multidimensional concept with both cognitive and emotional aspects that come into play as reusers search for, identify, and work with data that they did not originally collect.

In our DIPIR research, we examined data reusers’ trust in repositories and found that data reusers assess trust through repository functions—particularly data processing, metadata application, and data selection—and to a lesser extent repository actions, such as transparency.<sup>29</sup> These findings coincide with Adolfo

Prieto's work that identified clear repository policies, customer service, and systematic processes as increasing users' confidence in a repository's authenticity, integrity, and accessibility.<sup>30</sup> We also found that a repository's guarantee to preserve digital data and its overall reputation increased trust in the repository for certain disciplinary communities.<sup>31</sup>

In this chapter, we focus on trust in the data, because research suggests trust is “a key mediating variable between information quality and information usage.”<sup>32</sup> For example, Kelton, Fleischmann, and Wallace presented a general model of trust in information and contended that accuracy, objectivity, validity, and stability are important attributes leading to trust.<sup>33</sup> Donaldson and Conway confirmed these attributes but found that people considered authenticity, believability, coverage, currency, first-hand or primary nature, form, inaccurate information, and legibility important when assessing trust in archival documents.<sup>34</sup> Given these previous studies, we were interested in identifying which attributes data reusers might rely on to assess trust in data.

Across the three disciplines in the DIPIR study, we found that individuals used five trust markers when determining whether to reuse a dataset: the identity of the data producer, documentation, original peer-reviewed publications about the data, indications of prior reuse, and repository reputation (figure 4.1). Data reusers in each discipline mentioned data producers and documentation frequently, while only zoologists mentioned original peer-reviewed publications. Indications of prior reuse were primarily valued by social scientists, and repository reputation was important for both zoologists and social scientists.



**FIGURE 4.1**

Top five trust markers DIPIR study participants considered when assessing trust in data based on interviews with archaeologists ( $n=22$ ), social scientists ( $n=43$ ), and zoologists ( $n=27$ ).

## *Trust Marker: Data Producer*

Information about the data producer ranked highly as a trust marker across the three disciplines. Trust in the data producer was often mentioned in tandem with some other characteristic, such as the university where the research took place, the repository housing the data, or the university where the data producer trained. Archaeologist 13 provided this example:

Whose data do you trust? And it's primarily, it's sort of like, who do you know who does good work. So I go to the people from programs that are well known in the field, which for me is the Germans. People working out of handful of universities in Germany whether it's Munich or Tübingen, they are really well-taught. They know what they're doing, and they do it in a very standardized way.

## *Trust Marker: Documentation*

The level or quality of documentation for the data scored as another important indicator of trust across the disciplines. Reusers tended to focus on *how* the data were documented, rather than *what* was documented about the data. Characteristics of the documentation that were important included completeness or thoroughness of the record, evidence of standardized or professional practice, and the reuser's perception of its correctness. Zoologist 11 discussed how researchers' notebooks could reveal whether they were being systematic during data collection:

I used notebooks from multiple people... and some of them, through reading through them, I essentially did not fully trust the data they were collecting... I could tell they weren't doing it quite systematically enough.

## *Trust Marker: Publications and Prior Reuse Indicators*

Original peer-reviewed publications about the data were seen as an important indicator of trust for the zoologists, but much less so for social scientists and archaeologists. Zoologist 3 discussed using peer-reviewed literature to double-check information from museums:

For example, when I see records that look funny... like you know that's a mountain species, what would it be doing down there down by the sea, for example. I would then go to other published research about that group of species and see what people are saying if they... To try and cross-validate with the specimen data I'm using.

Indications of prior data reuse were most important for social scientists in assessing trust in the data. Social scientist 27 discussed prior reuse in reference to the Panel Study of Income Dynamics (PSID):

It has been around since 1968. It's heavily used. The dataset has been examined for problems like people dropping out of the study... And when you look at the PSID demographics there, since this was started in '68, how does that compare to a country now that has more immigrants, people where the demographics of the country has changed? Well, we've had a lot of work investigating all of these questions on PSID so these characteristics are pretty well known. It's an extremely trust-worthy dataset.

Prior reuse was less important for zoologists and not a factor for archaeologists. Prior reuse has become easier to track given the emerging data citation practices and the availability of alt.metrics showing downloads on some social science data repository websites.

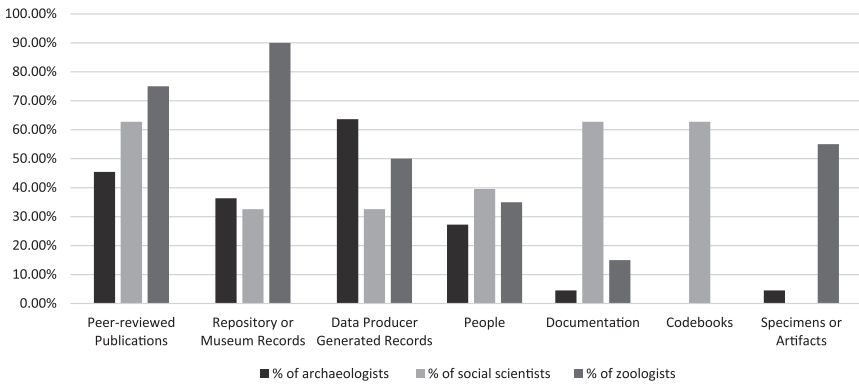
## *Trust Marker: Repository Reputation*

Both social scientists and zoologists ranked repository reputation highly. Members of these disciplines had experience with a wide variety of repositories and were able to differentiate reputations among them. The archaeologists in our study did not discuss repository reputation as a factor in assessing trust in the data. Social scientist 14 summed up the interplay of repository reputation and trust in the data as follows:

In general sort of I'm a lot more trustworthy of academic repositories and academically associated sources of data. Not that academics don't have political biases but they are subject to a lot of scrutiny in the academic community. So if you publish something with data that's clearly leaning towards one side you are going to get hammered or you are not going to get published really.

# Sources of Additional Support for Data Reuse

Prior research shows reusers often need different types of contextual information about data during the data reuse process and that they use a variety of sources to get it.<sup>35</sup> Our findings showed that the data reusers relied on seven key sources—peer-reviewed publications, repository or museum records, data producer-generated records, people, documentation, codebooks, and specimens or artifacts (figure 4.2). Some of these sources were employed across all three disciplines to different degrees and others were discipline-specific. The way in which researchers accessed and brought the sources together also varied given disciplinary and repository practices.



**FIGURE 4.2**

Seven key sources of contextual information that DIPIR study participants employed during the data reuse process based on interviews with archaeologists ( $n=22$ ), social scientists ( $n=43$ ), and zoologists ( $n=27$ ).

## Social Scientists

Social scientists primarily used codebooks, documentation, and peer-reviewed publications to facilitate data reuse. Our analysis showed the terms *codebook* and *documentation* were used interchangeably for this group, and they were the only group to specifically mention the term *codebook*. In order to create data records and documentation, ICPSR staff routinely used data collection and analysis information captured in the data producer-generated records. In addition, they added additional context such as missing data reports, descriptive statistics, data producer publications, and staff processing notes. By bringing all of this infor-

mation together, repository staff created one central point of access to much of the contextual information social scientists needed for data reuse. Social scientist 29 described some of the common types of contextual information social scientists sought in codebooks:

The codebook for me is more like; first, to get the label of the variable. Second, to get the meaning of the different categories that they have inside the dataset, and sometimes, yes they give you some basic descriptive statistics there, so you could see, “Okay, this [is] the range, this is the mean.”

ICPSR provides references to works citing the data held in its repository. Peer-reviewed publications were of particular interest to data reusers. Although the list was not exhaustive and the full text was not provided, social scientists were able to find publications that reference data of interest directly on the repository website. We found they used the publications in several ways during the data reuse process. For instance data producers’ publications were used to clarify data collection and analysis information, while data reusers’ publications were used to gauge the community interest and acceptance of the data. Social scientist 25 explained how peer-reviewed publications were used to discover data:

Yeah well, the story goes, when you research the literature you find out what datasets are to be used. Through the literature you find out what authors and investigators used to answer these questions. So you find out through into the literature....

Interestingly, even with long-existing repositories in place like the ICPSR, other people were a source of information for approximately 40 percent of the social scientists. For instance, social scientist colleagues provided opinions about reusing data and helped with data discovery. For novice data reusers like social scientist 09, professors provided reuse advice:

Because I’m so novice in these areas, I would heavily value the opinions of professors... even if I didn’t understand the reasons... I’m willing to accept that they know more about these areas than I do.

## *Archaeologists*

Like the social scientists, archaeologists relied on contextual information from data producer-generated records. Unlike the practices in social science, however,

the records were not assembled and repackaged into one source of information, like a codebook. Instead, archaeologists searched, reviewed, and assembled a number of data producer-generated records to support their reuse of data, such as geographic maps, stratigraphy drawings, tables of numerical data, images, artifact sketches or photos, field notes, and field reports. These records contained contextual information that was recorded in the field during an excavation or survey. For older studies, much of this information remained in paper-based form and could be accessed only in museums or through the data producer.

To a lesser extent than the social scientists and zoologists, archaeologists also relied on the peer-reviewed publications to facilitate data reuse. Archaeologists typically used data producer publications to discover and access data and additional contextual information since sharing and reusing archaeological data were relatively new phenomena. Archaeologists consulted people during the data reuse process, relying on museum staff and data producers primarily. These people resources were used in the same way as data producer publications: to discover and access data and contextual information. Sometimes, archeologists sought data producers' help through collaboration on data reuse studies as well. Archaeologist 09 described a visit to a museum to gather more data and to meet with the original excavator in order to clarify the contextual information found in several sources:

And so I started with the publications but I began to realize that I needed two things. One, I needed more data... I didn't have measurements on some of the artifacts and I needed that. And the other is I really needed to talk to the original excavator to find out some things that were confusing to me when I just looked at the photographs, or the maps, or the descriptions.... The materials were in Tulsa, Oklahoma. So I arranged to go to Tulsa and spend several days there getting the information I needed.... I sat down with the original excavator with maps and with everything else.... And he clarified a bunch of things for me....

## *Zoologists*

Most zoologists in our study mentioned using additional repository and museum records, followed by peer-reviewed literature, specimens, and data producer-generated records. All of these sources were used to access data collection and specimen information. The repository and museum records provided basic specimen information. Additionally, zoologists mentioned using the handwritten labels repository staff created when preparing specimens for preservation and photographs or x-rays of specimens if available.



The zoologists did not mention using peer-reviewed literature to gather information about prior reuse as frequently as social scientists, even though they commonly shared and reused zoological data. Instead zoologists used the literature, mainly journal articles, to discover and access data. Zoologist 07 talked about accessing data from a journal article that was “locked away” in a pdf format:

The literature I’m using here is there are tables of fossil localities or presence in the taxa used for various analyses and those can prove to be useful. So I’m scraping, you know, I’m actually taking content that is locked away in pdf and converting that into a format so I can reuse for analysis of the data.

Zoologists also mentioned using physical specimens much more than archaeologists used the physical artifacts. Several zoologists discussed visiting museums or requesting that physical specimens be sent to them to gather additional sequence or morphometric (e.g., size, shape, color, etc.) data. Others requested physical specimens to verify identification of species. Zoologist 19 discussed the desire to examine the voucher (i.e., representative) specimen from which DNA was drawn:

If somebody misidentifies the fish, or the tree or whatever it is you’re looking at and uploads it to GenBank with an incorrect taxon identifier, that causes downstream problems, particularly if they didn’t save a voucher so that someone can verify the ID. So when I do reuse specimens, I try to get a photograph of the voucher, or actually look at the true voucher itself, to verify that the original person that deposited the DNA sequence had correctly identified the species from which it was taken.

Zoologists also reported using data producer-generated records, but unlike archaeologists, they used fewer types, field notebooks primarily, to access the contextual information captured during data collection. Zoologists did not generate as many records during data collection. Zoologists also relied on people, collection managers at museums and data producers, to get additional contextual information. Collection managers were called on prior to a museum visit to get more information about a collection and to make arrangements to ensure a worthwhile visit as well.

The sources of support researchers in our study mentioned using depended primarily on how the data were documented in and out of the field. In both social science and zoology, there were dedicated repository and museum staff with expertise in particular types of collections that could help data producers manage and curate the data and the associated documentation. The same was not true

for archaeology. Fewer dedicated staff were available to help. These differences influenced whether and how documentation about the data were represented and disseminated via repositories or other information sources.

## Implications for Repository Practice

Data sharing and reuse are increasing and will continue to do so for the foreseeable future. Our chapter aims to provide insight into the needs of data reusers, knowing that disciplinary practice is not always aligned with the changes afoot. It takes time. Not only do disciplinary practices need to change, but repository infrastructures need to mature. Both are well underway, but we believe that a broader understanding of the designated community of users, particularly data reusers, is needed. In the paragraphs that follow we discuss the implications our findings have on repository practice.

We found that the repository and museum staff within the social science and zoology disciplines play a key role in readying data for reuse. They manage, prepare, and curate data for deposit—steps that allow them to create central access points to data and other sources of contextual information that data reusers may need. The staff's work also lightens the data producers' load, especially when the staff intervene at the beginning of the data life cycle and can negotiate curation goals and needs of the data producers, repository staff, and data reusers concurrently. This level of support is beginning to happen within the archaeology discipline as well. Open Context staff is working on a project to intervene during the data collection process at the archaeological site. Staff plans to examine data producers' practices during excavations in order to provide guidance on recording and managing data in ways that make repository staff's downstream activities easier and better align data creation practices with meaningful reuse.<sup>36</sup> Similar to each of these three communities, we suggest that all repository staff find ways to center themselves within their designated community of users to better understand upstream and downstream needs in order to align them with repository staff's data deposit and curation activities.

We also found that reusers' trust in data is not informed only through their encounters with the data. Reusers rely on a variety of factors at play during the data life cycle, such as how the data are documented, where the data producers were trained, what university they represented when the research took place, and the trustworthiness of the repository housing the data. These characteristics are not always captured and disseminated through a repository. Furthermore, they are signifiers for reputational perceptions and opinions that get formed over time as one gains experience within the discipline, with the data, and with the repository.

We suggest that repository staff make themselves aware of these trust indicators and consider ways to more readily shape reusers' opinions about the data being offered from the repository. Repository reputation and documentation quality, in particular, can be shaped by repository staff to meet reusers' expectations.

Our findings indicate that all repositories do not have to house all of the contextual information associated with the data to be effective. However, they do have to provide data reusers access to provenance information and pointers to the additional contextual information about the data if housed elsewhere. Take Genbank, GBIF, and VertNet as examples. As repositories that aggregate zoological data across museums to facilitate easier search and discovery of species, they upload some, but not all of the contextual information associated with specimens. The museums where the data were originally deposited and the rich metadata originally created remain responsible for managing, curating, and preserving the data and contextual information. In this case, repository staff at multiple institutions made an informed decision about data stewardship and data services, given the needs of their respective data producers and reusers and the repository infrastructures in place at each other's institutions. We suggest other repository staff do the same. Consider the types of partnerships that can be formed with other repositories to complement and extend each other's capabilities and to add value to the designated community of users.

Regardless of the growth in repositories or how well established they have become, data producers remain an important source of contextual information for data reusers. By reaching out to and developing relationships with the data producers, repository staff can provide reusers with another way to learn about the data. By monitoring these engagements, repository staff also can benefit by understanding the unmet needs of their designated community of users and adapt accordingly, hopefully reducing its reliance on data producers' memories over time. Knowing that novice data reusers sought advice from expert data reusers, repository staff might want to talk to both groups to determine whether there are user interface design changes, instructional modules, or other scaffolding that the repository can provide to improve the novice data reusers' experience.

Our findings show that data reusers across the three disciplines supplement their data reuse with peer-reviewed publications. Archaeologists and zoologists rely on data producer publications primarily, whereas social scientists rely on data reuser publications. The differences are likely due to the maturity of data sharing and reuse within the disciplines. Archaeologists and zoologists are using data producers' publications to access data and contextual information, whereas social scientists are using data reusers' publications to gauge the social science community's interest and acceptance of data for reuse. We liken the latter to a dataset peer review, of sorts. In disciplines where data sharing and reuse are still in the early stages and peer-reviewed publications are limited, repository staff might consider assembling a team of experts to provide a peer review of the data. We

also suggest that repository staff capture data reuse metrics, provide DOIs and suggestions for data citation, and maintain bibliographies of data reuse studies for researchers to incorporate into their decisions to trust and reuse the data.

## Conclusion

Disciplinary practices and traditions are the guiding forces in the development of a data sharing and reuse culture. Yet external forces, such as technological advances, federal mandates and policies, and repository infrastructure have shaped them further. We examined data reuse, a less-studied phenomenon, because we believe that the knowledge held by the designated communities of users, particularly re-users, is needed. We see it as a way to further develop repository infrastructure in ways that will align with the cultural changes needed in many disciplines to make data reuse a valued and viable alternative or supplement to original data collection.

## Acknowledgments

The DIPIR Project was made possible by a National Leadership Grant from the Institute of Museum and Library Services, LG-06-10-0140-10, “Dissemination Information Packages for Information Reuse,” and support from OCLC, and University of Michigan. We thank members of the DIPIR team, including University of Michigan students, research Fellows, institutional partners, and individual collaborators. We also thank manuscript reviewers and editors for their insightful comments and suggestions.

## Notes

1. Morgan Daniels, Ixchel Faniel, Kathleen Fear, and Elizabeth Yakel, “Managing Fixity and Fluidity in Data Repositories,” in *Proceedings of the 2012 iConference* (Toronto: ACM, 2012), 279–86, doi:10.1145/2132176.2132212; Rebecca D. Frank, Adam Kriesberg, Elizabeth Yakel, and Ixchel M. Faniel, “Looting Hoards of Gold and Poaching Spotted Owls: Data Confidentiality Among Archaeologists and Zoologists,” in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community* (Silver Springs, MD: American Society for Information Science, 2015), 37:1–37:10, <http://dl.acm.org/citation.cfm?id=2857070.2857107>; Rebecca D. Frank, Elizabeth Yakel, and Ixchel M. Faniel, “Destruction/Reconstruction: Preservation of Archaeological and Zoological Research Data,” *Archival Science* 15, no. 2 (January 11, 2015): 141–67, doi:10.1007/s10502-014-9238-9; Adam Kriesberg, Rebecca D. Frank, Ixchel M. Faniel, and Elizabeth Yakel, “The Role of Data Reuse in the Apprenticeship Process,” *Proceedings of the American Society for Information Science and Technology* 50, no. 1 (2013): 1–10, doi:10.1002/meet.14505001051.

2. Janice M. Morse and Linda Niehaus, *Mixed Method Design* (Walnut Creek, CA: Left Coast Press, 2009).
3. Inter-university Consortium for Political and Social Research, "ICPSR Bibliography of Data-Related Literature," May 2014, <http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/>.
4. John W. Creswell, *Research Design* (Los Angeles: Sage, 2009).
5. Tony Hey and Anne Trefethen, "The Data Deluge: An E-Science Perspective," in *Grid Computing*, ed. Fran Berman, Geoffrey Fox, and Tony Hey (Hoboken, NJ: John Wiley & Sons, 2003), 809–24.
6. Office of Management and Budget, "Circular A-110 Revised 11/19/93 As Further Amended 9/30/99," White House website, last updated September 30, 1999, [http://www.whitehouse.gov/omb/circulars\\_a110](http://www.whitehouse.gov/omb/circulars_a110).
7. John P. Holdren, "Increasing Access to the Results of Federally Funded Scientific Research," Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013, [https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
8. Tyler Walters and Judy Ruttenberg, "SHared Access Research Ecosystem," *Educause Review* 49, no. 2 (2014): 56–57.
9. Benjamin S. Arbuckle, Sarah Whitcher Kansa, Eric Kansa, David Orton, Canan Çakırlar, Lionel Gourichon, Levent Atici, et al., "Data Sharing Reveals Complexity in the Westward Spread of Domestic Animals across Neolithic Turkey," ed. Stephen Shennan, *PLoS ONE* 9, no. 6 (June 13, 2014): e99845, doi:10.1371/journal.pone.0099845.
10. Myron P. Gutmann, Mark Abrahamson, Margaret O. Adams, Micah Altman, Caroline Arms, Kenneth Bollen, Michael Carlson, et al., "From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data," *Library Trends* 57, no. 3 (2009): 315–37, doi:10.1353/lib.0.0039.
11. Kalpana Shankar, Kristin Eschenfelder, and Greg Downey, "Studying the History of Social Science Data Archives as Knowledge Infrastructure," *Science and Technology Studies* (in press).
12. Mary Vardigan, Pascal Heus, and Wendy Thomas, "Data Documentation Initiative: Toward a Standard for the Social Sciences," *International Journal of Digital Curation* 3, no. 1 (August 6, 2008): 107–13, doi:10.2218/ijdc.v3i1.45.
13. Ixchel M. Faniel, Adam Kriesberg, and Elizabeth Yakel, "Social Scientists' Satisfaction with Data Reuse," *Journal of the Association for Information Science and Technology* 67, no. 6 (June 2015): 1404–16, doi:10.1002/asi.23480.
14. Sarah Whitcher Kansa and Eric Kansa, "Beyond Bone Commons: Recent Developments in Zooarchaeological Data Sharing," *SAA Archaeological Record* 11, no. 1 (January 2011): 26–29.
15. Eric C. Kansa, Jason Schultz, and Bissell N. Ahrash, "Protecting Traditional Knowledge and Expanding Access to Scientific Data: Juxtaposing Intellectual Property Agendas via a 'Some Rights Reserved' Model," *International Journal of Cultural Property*, no. 12 (2005): 285–314, doi:10.1017/S0940739105050204.
16. Society of American Archaeology, *Editorial Policy, Information for Authors, and Style Guide for American Antiquity, Latin American Antiquity and Advances in Archaeological Practice* (Washington, DC: Society of American Archaeology, last revised October 14, 2014), 5, [http://www.saa.org/portals/0/saa/publications/styleguide/styleguide\\_final\\_813.pdf](http://www.saa.org/portals/0/saa/publications/styleguide/styleguide_final_813.pdf).

17. Frank, Yakel, and Faniel, "Destruction/Reconstruction."
18. Eric C. Kansa and Sarah Whitcher Kansa, "We All Know That a 14 Is a Sheep: Data Publication and Professionalism in Archaeological Communication," *Journal of Eastern Mediterranean Archaeology and Heritage Studies*, Project Muse, 1, no. 1 (March 16, 2013): 88–97.
19. Ixchel Faniel, Eric Kansa, Sarah Whitcher Kansa, Julianna Barrera-Gomez, and Elizabeth Yakel, "The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse," in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, (New York: ACM, 2013), 295–304, doi:10.1145/2467696.2467712.
20. Arbuckle et al., "Data Sharing Reveals Complexity."
21. Levent Atici, Sarah Kansa, Justin Lev-Tov, and Eric Kansa, "Other People's Data: A Demonstration of the Imperative of Publishing Primary Data," *Journal of Archaeological Method and Theory*, April 11, 2012, 1–19, doi:10.1007/s10816-012-9132-9.
22. Diane Harley, Sophia Krzys Acord, Sarah Earl-Novell, Shannon Lawrence, and C. Judson King, *Assessing the Future Landscape of Scholarly Communication* (Berkeley: University of California Press, 2010).
23. Faniel et al., "The Challenges of Digging Data"; Kansa and Kansa, "We All Know That a 14 Is a Sheep."
24. Stephen T. Asma, *Stuffed Animals and Pickled Heads* (Oxford; New York: Oxford University Press, 2003); Susan Leigh Star and James R. Griesemer, "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39," *Social Studies of Science* 19, no. 3 (August 1, 1989): 387–420, doi:10.1177/030631289019003001.
25. Star and Griesemer, "Institutional Ecology, 'Translations' and Boundary Objects."
26. University of California Berkeley, Museum of Vertebrate Zoology, "MVZ Guidelines for Scientific Collecting Permits," accessed March 18, 2016, [http://mvz.berkeley.edu/Sci\\_Coll\\_Permits.html](http://mvz.berkeley.edu/Sci_Coll_Permits.html). This website provides a good overview of the regulatory environment zoologists face when collecting specimens in the field.
27. Ayoung Yoon, "End Users' Trust in Data Repositories: Definition and Influences on Trust Development," *Archival Science* 14, no. 1 (March 2014): 19, doi:10.1007/s10502-013-9207-8.
28. Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer, "Not So Different after All: A Cross-Discipline View of Trust," *Academy of Management Review* 23, no. 3 (July 1, 1998): 395.
29. Elizabeth Yakel, Ixchel Faniel, Adam Kriesberg, and Ayoung Yoon, "Trust in Digital Repositories," *International Journal of Digital Curation* 8, no. 1 (June 14, 2013): 143–56, doi:10.2218/ijdc.v8i1.251.
30. Adolfo G. Prieto, "From Conceptual to Perceptual Reality: Trust in Digital Repositories," *Library Review* 58, no. 8 (2009): 593–606, doi:10.1108/00242530910987082.
31. Yakel et al., "Trust in Digital Repositories."
32. Kari Kelton, Kenneth R Fleischmann, and William A Wallace, "Trust in Digital Information," *Journal of the American Society for Information Science and Technology* 59, no. 3 (February 1, 2008): 363, doi:10.1002/asi.20722.
33. Ibid.
34. Devan Ray Donaldson and Paul Conway, "User Conceptions of Trustworthiness for Digital Archival Documents: User Conceptions of Trustworthiness for Digital Archival Documents," *Journal of the Association for Information Science and Technology* 66, no. 12 (2015): 2427–44, doi:10.1002/asi.23330.

35. I. M. Faniel and T. E. Jacobsen, "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data," *Computer Supported Cooperative Work* 19, no. 3–4 (August 2010): 355–75, doi:10.1007/s10606-010-9117-8; Betsy Rolland and Charlotte P. Lee, "Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research," in *CSCW '13: Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, TX: ACM, 2013), 435–44; Ann Zimmerman, "Not by Metadata Alone: The Use of Diverse Forms of Knowledge to Locate Data for Reuse," *International Journal on Digital Libraries* 7, no. 1–2 (2007): 5–16, doi:10.1007/s00799-007-0015-8; Faniel et al., "The Challenges of Digging Data"; Kriesberg et al., "The Role of Data Reuse in the Apprenticeship Process"; Ixchel M. Faniel, Adam Kriesberg, and Elizabeth Yakel, "Data Reuse and Sensemaking among Novice Social Scientists," *Proceedings of the American Society for Information Science and Technology* 49, no. 1 (2012): 1–10, doi:10.1002/meet.14504901068.
36. Alexandria Archive Institute, "Bridging Data Creation and Reuse," accessed March 8, 2016, <http://alexandriaarchive.org/projects/bridging-creation-and-reuse/>.

## Bibliography

- Alexandria Archive Institute. "Bridging Data Creation and Reuse." Accessed March 8, 2016. <http://alexandriaarchive.org/projects/bridging-creation-and-reuse/>.
- Arbuckle, Benjamin S., Sarah Whitcher Kansa, Eric Kansa, David Orton, Canan Çakırlar, Lionel Gourichon, Levent Atici, et al. "Data Sharing Reveals Complexity in the Westward Spread of Domestic Animals across Neolithic Turkey." Edited by Stephen Shennan. *PLoS ONE* 9, no. 6 (June 13, 2014): e99845. doi:10.1371/journal.pone.0099845.
- Asma, Stephen T. *Stuffed Animals and Pickled Heads: The Culture and Evolution of Natural History Museums*. Oxford; New York: Oxford University Press, 2003.
- Atici, Levent, Sarah Kansa, Justin Lev-Tov, and Eric Kansa. "Other People's Data: A Demonstration of the Imperative of Publishing Primary Data." *Journal of Archaeological Method and Theory*, April 11, 2012, 1–19. doi:10.1007/s10816-012-9132-9.
- Creswell, John W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Los Angeles: Sage, 2009.
- Daniels, Morgan, Ixchel Faniel, Kathleen Fear, and Elizabeth Yakel. "Managing Fixity and Fluidity in Data Repositories." In *Proceedings of the 2012 iConference*, 279–86. Toronto: ACM, 2012. doi:10.1145/2132176.2132212.
- Donaldson, Devan Ray, and Paul Conway. "User Conceptions of Trustworthiness for Digital Archival Documents." *Journal of the Association for Information Science and Technology* 66, no. 12 (2015): 2427–44. doi:10.1002/asi.23330.
- Faniel, I. M., and T. E. Jacobsen. "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data." *Computer Supported Cooperative Work* 19, no. 3–4 (August 2010): 355–75. doi:10.1007/s10606-010-9117-8.
- Faniel, Ixchel, Eric Kansa, Sarah Whitcher Kansa, Julianna Barrera-Gomez, and Elizabeth Yakel. "The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse." In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 295–304. New York: ACM, 2013. doi:10.1145/2467696.2467712.



- Faniel, Ixchel M., Adam Kriesberg, and Elizabeth Yakel. "Data Reuse and Sensemaking among Novice Social Scientists." *Proceedings of the American Society for Information Science and Technology* 49, no. 1 (2012): 1–10. doi:10.1002/meet.14504901068.
- . "Social Scientists' Satisfaction with Data Reuse." *Journal of the Association for Information Science and Technology* 67, no. 6 (June 2015): 1404–16. doi:10.1002/asi.23480.
- Frank, Rebecca D., Adam Kriesberg, Elizabeth Yakel, and Ixchel M. Faniel. "Looting Hoards of Gold and Poaching Spotted Owls: Data Confidentiality among Archaeologists and Zoologists." In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, 37:1–37:10. Silver Springs, MD: American Society for Information Science, 2015. <http://dl.acm.org/citation.cfm?id=2857070>.[2857107](http://dl.acm.org/citation.cfm?id=2857070)[http://dl.acm.org/citation.cfm?id=2857070.2857107](http://dl.acm.org/citation.cfm?id=2857070).
- Frank, Rebecca D., Elizabeth Yakel, and Ixchel M. Faniel. "Destruction/Reconstruction: Preservation of Archaeological and Zoological Research Data." *Archival Science* 15, no. 2 (January 11, 2015): 141–67. doi:10.1007/s10502-014-9238-9.
- Gutmann, Myron P., Mark Abrahamson, Margaret O. Adams, Micah Altman, Caroline Arms, Kenneth Bollen, Michael Carlson, et al. "From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data." *Library Trends* 57, no. 3 (2009): 315–37. doi:10.1353/lib.0.0039.
- Harley, Diane, Sophia Krzys Acord, Sarah Earl-Novell, Shannon Lawrence, and C. Judson King. *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines*. Berkeley: University of California Press, 2010.
- Hey, Tony, and Anne Trefethen. "The Data Deluge: An E-Science Perspective." In *Grid Computing*, edited by Fran Berman, Geoffrey Fox, and Tony Hey, 809–24. Hoboken, NJ: John Wiley & Sons, 2003.
- Holdren, John P. "Increasing Access to the Results of Federally Funded Scientific Research." Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013. [https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
- Inter-university Consortium for Political and Social Research. "ICPSR Bibliography of Data-Related Literature." May 2014. <http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/>.
- Kansa, Eric C., and Sarah Whitcher Kansa. "We All Know That a 14 Is a Sheep: Data Publication and Professionalism in Archaeological Communication." *Journal of Eastern Mediterranean Archaeology and Heritage Studies*, Project Muse, 1, no. 1 (March 16, 2013): 88–97.
- Kansa, Eric C., Jason Schultz, and Bissell N. Ahrash. "Protecting Traditional Knowledge and Expanding Access to Scientific Data: Juxtaposing Intellectual Property Agendas via a 'Some Rights Reserved' Model." *International Journal of Cultural Property*, no. 12 (2005): 285–314. doi:10.1017/S0940739105050204.
- Kansa, Sarah Whitcher, and Eric Kansa. "Beyond BoneCommons: Recent Developments in Zooarchaeological Data Sharing." *SAA Archaeological Record* 11, no. 1 (January 2011): 26–29.
- Kelton, Kari, Kenneth R. Fleischmann, and William A. Wallace. "Trust in Digital Information." *Journal of the American Society for Information Science and Technology* 59, no. 3 (February 1, 2008): 363–74. doi:10.1002/asi.20722.



- Kriesberg, Adam, Rebecca D. Frank, Ixchel M. Faniel, and Elizabeth Yakel. "The Role of Data Reuse in the Apprenticeship Process." *Proceedings of the American Society for Information Science and Technology* 50, no. 1 (2013): 1–10. doi:10.1002/meet.14505001051.
- Morse, Janice M., and Linda Niehaus. *Mixed Method Design: Principles and Procedures*. Walnut Creek, CA: Left Coast Press, 2009.
- Office of Management and Budget. "Circular A-110 Revised 11/19/93 As Further Amended 9/30/99." White House website. Last updated September 30, 1999. [http://www.whitehouse.gov/omb/circulars\\_a110](http://www.whitehouse.gov/omb/circulars_a110).
- Prieto, Adolfo G. "From Conceptual to Perceptual Reality: Trust in Digital Repositories." *Library Review* 58, no. 8 (2009): 593–606. doi:10.1108/00242530910987082.
- Rolland, Betsy, and Charlotte P. Lee. "Beyond Trust and Reliability: Reusing Data in Collaborative Cancer Epidemiology Research." In *CSCW '13: Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 435–44. San Antonio, TX: ACM, 2013.
- Rousseau, Denise M., Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. "Not So Different after All: A Cross-Discipline View of Trust." *Academy of Management Review* 23, no. 3 (July 1, 1998): 393–404.
- Shankar, Kalpana, Kristin Eschenfelder, and Greg Downey. "Studying the History of Social Science Data Archives as Knowledge Infrastructure." *Science and Technology Studies* (in press).
- Society of American Archaeology. *Editorial Policy, Information for Authors, and Style Guide for American Antiquity, Latin American Antiquity and Advances in Archaeological Practice*. Washington, DC: Society of American Archaeology, last revised October 14, 2014. [http://www.saa.org/portals/0/saa/publications/styleguide/styleguide\\_final\\_813.pdf](http://www.saa.org/portals/0/saa/publications/styleguide/styleguide_final_813.pdf).
- Star, Susan Leigh, and James R. Griesemer. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39." *Social Studies of Science* 19, no. 3 (August 1, 1989): 387–420. doi:10.1177/030631289019003001.
- University of California Berkeley, Museum of Vertebrate Zoology. "MVZ Guidelines for Scientific Collecting Permits." Accessed March 18, 2016. [http://mvz.berkeley.edu/Sci\\_Coll\\_Permits.html](http://mvz.berkeley.edu/Sci_Coll_Permits.html).
- Vardigan, Mary, Pascal Heus, and Wendy Thomas. "Data Documentation Initiative: Toward a Standard for the Social Sciences." *International Journal of Digital Curation* 3, no. 1 (August 6, 2008): 107–13. doi:10.2218/ijdc.v3i1.45.
- Walters, Tyler, and Judy Ruttenberg. "SHared Access Research Ecosystem." *Educause Review* 49, no. 2 (2014): 56–57.
- Yakel, Elizabeth, Ixchel Faniel, Adam Kriesberg, and Ayoung Yoon. "Trust in Digital Repositories." *International Journal of Digital Curation* 8, no. 1 (June 14, 2013): 143–56. doi:10.2218/ijdc.v8i1.251.
- Yoon, Ayoung. "End Users' Trust in Data Repositories: Definition and Influences on Trust Development." *Archival Science* 14, no. 1 (March 2014): 17–34. doi:10.1007/s10502-013-9207-8.
- Zimmerman, Ann. "Not by Metadata Alone: The Use of Diverse Forms of Knowledge to Locate Data for Reuse." *International Journal on Digital Libraries* 7, no. 1–2 (2007): 5–16. doi:10.1007/s00799-007-0015-8.





## CHAPTER 5\*

# Overlooked and Overrated Data Sharing

## Why Some Scientists Are Confused and/or Dismissive

*Heidi J. Imker*

Data curation, particularly within academic libraries, has gained appreciable momentum by developing an energetic community dedicated to providing widespread access to well-curated data. In one vision of the future, the data required to validate or extend a research study is readily available, and the publication of data itself will bear an equal importance to that of the article publication. The data curation community is eager to help catalyze that transformation through services and advocacy. Yet in practice, it's not uncommon to encounter scientists who question the cost-benefit ratio of the time and effort involved with curation, publication, and preservation of research data. How can something that seems so self-evident to the data curation community be so challenging to implement in the wild?

One possible reason is that libraries and the data curation community gravitate towards the progressive ideals of open science;<sup>1</sup> however, by its very nature, progressive is not representative. Data curators are well acquainted with the shortcomings of current data sharing practices, such as the over-use of PDFs for data publication, which restricts reuse by encapsulating otherwise useful data in this traditional publication format. However, such practices have been in place

---

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

for decades, and frustration with those practices is not uniform; there is rarely one voice that emerges from a given community of practice, let alone unification across all research communities.<sup>2</sup> The aim of this chapter is to take a fresh look at current practices and the nuances that surround data sharing in order to hone our messages and services as data curators with a range of perspectives in mind.

This chapter will first contextualize data sharing in the United States by looking at cultural expectations and norms within science communities. We'll then examine how scientists have historically shared research data, particularly long before modern public access requirements, since this is a useful way to frame current practice. Overlooking presently active, albeit seemingly imperfect, forms of data sharing, while ignoring researchers' own experiences and perspectives, can lead to confused or dismissive reactions to data sharing mandates and outreach. Understanding this challenge is key for those in the data curation community who are attempting to garner researcher buy-in for resources and services in support data sharing activities. In particular, some forms of sharing are successful and worthy of reexamination in light of their prevalence and adoption, even if they involve methods that do not meet data curation community approval. Finally, several large-scale data sharing efforts have been unsuccessful, and examination of the circumstances that led to their sun-setting is informative as well. The data curation community is understandably receptive to the issues that drive increased data sharing, namely transparency, reuse, and reproducibility, but we must also acknowledge the limitations of data sharing for the healthy and sustainable development of the data curation field.

## Data Sharing in Context

As funding shrinks and expectations expand, it is not surprising that researchers consistently list time, cost, and appropriateness (such as sensitivity, confidentiality, or IP protection) as barriers to data sharing.<sup>3</sup> In 2005, the administrative burdens required to execute federally funded research became so overwhelming and problematic that the topic escalated to large-scale review by the Federal Demonstration Partnership.<sup>4</sup> Despite some efforts to reform and streamline reporting activities over the following decade, only 57.7 percent of faculty's available research time was actually spent on active research.<sup>5</sup> The rest of the time was spent on administrative tasks for research, largely preparing new proposals and reporting on awarded grants. While data-sharing efforts could be considered as part of active research, it cannot be ignored that the time available for *all* aspects of active research is limited.

The need for extramural funding in the sciences feeds directly into the time shortage mentioned above. The percentage of US grants submitted that are successfully awarded, known as "funding success," has steadily decreased in recent

years, from roughly 1 in 3 being awarded in 2001 to roughly 1 in 5 being awarded in 2013.<sup>6</sup> Reduction in funding success can be attributed to many causes, but both increased demand (i.e., more grant proposals submitted) and less federal funding when adjusted for inflation are prominent reasons.<sup>7</sup> Loss of grant funding in the sciences, especially over extended periods of time, results in the inability to fund material purchase, equipment allocation, and graduate student, postdoc, or staff salaries. This dramatically slows project progression, including publication activity, and reduces subsequent competitiveness on future applications. For example, when an investor submits an application to renew an NIH grant, the review panel “will consider the progress made in the last funding period,” and the criteria include demonstration of “an ongoing record of accomplishments that have advanced their field(s).”<sup>8</sup> When Tenopir’s 2015 follow-up survey on data-sharing practices and perceptions included “I need to publish first” as a potential barrier, it was rendered the new top concern through affirmation by 43.5 percent of respondents.<sup>9</sup>

Grants are also a source of support for institutions through recovery of operating costs (e.g., administrative support, operation and maintenance of physical space, etc.). Recovery occurs by application of an “indirect cost rate” to funds awarded, and the rate is derived through a negotiation between the grantee institution and funding organization. Rates may vary dramatically, but for illustrative purposes we can use an average rate of 58.2 percent based on 49 institutional rates recently compiled.<sup>10</sup> In the most straightforward scenario, if an investigator is awarded \$100,000 in direct costs for a project, an additional \$58,200 is provided to the institution for indirect costs, resulting in a total award of \$158,200 from the funding agency. Thus fewer grants mean less funding not just for the investigator, but also for the institution.

With productivity hampered and financial pressure at the institution, loss of funding for a faculty member may come with marginalization within the scientific community and within the institution. Marginalization at the institution may result in reduction in lab space, increased teaching or administrative load, or lack of input into decisions. Tenured faculty are by no means immune to marginalization, but a lack of funding for untenured faculty places them at a distinct disadvantage. As a result, in the sciences pretenure faculty are urged to focus on securing external grants as a requirement for promotion.<sup>11</sup> While cultural changes for openness and sharing may be occurring, the reality for today—and most likely for several years to come—is that the average academic scientist will focus his or her finite time on what ensures continued funding and job security. And as data curators we must think strategically to work within this reality.

Therefore, as we consider data curation work, it’s important to keep in mind that a single definition of what constitutes data sharing cannot be extrapolated across all domains, since scientific disciplines themselves have the latitude to define what data means within each of their disciplines.<sup>12</sup> In fact, even within

domains, data sharing takes on an a myriad of forms; for example, the U.S. Geological Survey Manual states “USGS scientific data may be released or disseminated in a variety of ways, for example in datasets and databases, software, and other information products including USGS series publications (SM 1100.3), outside publications (SM 1100.4), and USGS Web pages.”<sup>13</sup> This sort of cultural relativism may be a frustration within the data curation community since it could possibly enable data withholding. However, disciplines are grappling with the current ambiguity of “data” itself,<sup>14</sup> let alone “data sharing.”<sup>15</sup> This isn’t entirely surprising. During examination of a similar semantic data topic, Renear, Sacchi, and Wickett stated that while a precise definition of *dataset* is desirable to the data curation community, informational definitions are generally functional and specific to a given discipline.<sup>16</sup> Efforts to define data sharing on behalf of a community are likely to be dismissed, and by talking at cross-purposes, data curators may lose the opportunity to nurture the evolution of those definitions within scientific communities.

While the data curation community often focuses on scientists not sharing research data, evidence that scientists do share data is prolific. Many reports, including surveys, case studies, and even data-withholding studies, indicate successful data sharing does exist. For example, surveys of researcher data-sharing practices consistently report that researchers do share their data. In 2011 Tenopir and colleagues found that only 9.6 percent of respondents somewhat or strongly disagreed with the statement “I share my data with others,” whereas the vast majority of respondents, 74.9 percent, strongly or somewhat agreed with the statement; the majority believed that they were sharing data at least to some extent.<sup>17</sup> Moreover, Tenopir and colleagues found that this sentiment increased in the 2015 follow-up study.<sup>18</sup> The 2014 Wiley study on data sharing found that 36 percent to 66 percent of researchers across five major disciplines self-reported sharing their data.<sup>19</sup> Within this study, the highest reported reason for hesitancy was intellectual property or confidentiality issues, both of which are well-acknowledged exceptions, even within the OSTP memo itself.<sup>20</sup> These concerns may account for the discipline reporting the lowest sharing: social scientists; however, openly shared data for the Wiley survey is ironically not yet available. A few empirical studies of data withholding have shown less data sharing in practice than the self-reporting survey results, albeit several of these studies have also been in disciplines that involve human subject research and therefore are more likely to be subject to ethical concerns.<sup>21</sup> Regardless of sensitivities, the results did not conclude that zero data sharing occurred. Examination of articles postpublication for evidence of shared data also revealed that sharing routinely occurs in practice and is not just an unsought ideal.<sup>22</sup>

Although not the focus of this chapter, it’s important to note the seemingly conflicting messages being directed at researchers regarding sensitive data. In particular, the rigorous procedures required for protecting human subjects car-

ry serious ramifications if breached, and researchers are constantly reminded of their obligations.<sup>23</sup> Furthermore, when the White House announced policies for government-generated Open Data, it also warned of individual identification through the “mosaic effect,” which occurs when nonidentifying data is combined with other available data to enable identification.<sup>24</sup> While a supposed fear of inappropriate disclosure could be used as a crutch to avoid data sharing, in this complex environment one person’s data withholding may be another person’s genuine concern about data breach or lack of adequate informed consent. Social and behavioral sciences have developed methodologies, protocols, and systems to allow appropriate dissemination of some restricted-use data, and repositories such as ICPSR offer excellent resources and guidance.<sup>25</sup> Thoughtful implementation of appropriate procedures and practices must be crafted at the point of project conceptualization such that the results are ultimately useful to the research community but also safe and ethical for participants. Through proactive engagement with researchers, data curators can be the gateway to such information before a study even begins and therefore increase the likelihood that study design will enable future data sharing.

Given this environment, how have scientists traditionally shared data? The next sections of this chapter will explore several overlooked ways in which researchers may already be sharing their data.

## *Overlooked Data Sharing: Article Publication*

Scientists frequently think of article publication as a form of data sharing, and it is critical to acknowledge not only that this concept exists, but also that it has been recapitulated throughout their communities, including funding agencies. As of October 2015, NIH’s Data Sharing workbook still says “Some studies, such as small laboratory-based projects, make raw data available in publications.”<sup>26</sup> Likewise, example data management guidance available from NSF and USGS websites reference data sharing via publication.<sup>27</sup> In an analysis of 1,260 Data Management Plans (DMPs) submitted for NSF applications at the University of Illinois, Mischo, Schlembach, and O’Donnell found “publication” listed as a data sharing mechanism 44 percent of the time.<sup>28</sup> So herein lies an important cultural disconnect in data sharing: as data curators, we are overlooking what many in scientific communities believe is an acceptable form of data sharing because it doesn’t fit into *our* definition of data sharing. It cannot be overemphasized that what may be substandard for the data curation community does not trump what is standard for a community of practice; *cultural norms are a critical driver for practice.*<sup>29</sup>

As an analogy, let's consider an example where someone uploads a presentation to a web service, where the slide deck is saved as a PDF without comments, animation, or the ability to manipulate. While it's obvious the slides could be shared in a manner more amenable to reuse by providing the original presentation format, could one say that the person who posted slides via PDF did not share because the format precludes ready reuse? Is sharing in this context really a true-or-false question? It might be worth reinforcing that accuracy is fundamental to science, and therefore the question of data sharing itself is confusing when presented through application of Boolean logic, with binary true/false variables. Fuzzy logic, with many-valued variables, is more appropriate. For that reason, our messages to scientists must emphasize *how* data is shared as opposed to the singular act of data sharing itself. Amending and clarifying our language by using phrases such as "reuse-ready sharing," "fit-for-purpose sharing," or "source file sharing" is one step in that direction.

Consider a recent study from Ron Vale that examined the amount of data shared through publication by comparing figures in publications in the journals *Nature*, *Cell*, and the *Journal of Biological Chemistry* for years 1984 and 2014.<sup>30</sup> Figures are a critical component of academic work and can present data (including raw, aggregate, and representative) through tables, graphs, images, schematics, and more. Through scoring of figures and panels, Vale concluded that publications include 2 to 4 times more data in 2014 than they did in 1984. The increase in data-per-publication ties into time-to-publication, which has slowed according to Vale's analysis. He attributed both trends largely to the need to publish comprehensive studies that provide an exhaustive and, especially in the eyes of the reviewer, hopefully unequivocal argument that the findings are valid. This sentiment has been echoed elsewhere during interviews with scientists.<sup>31</sup> Interestingly, Vale expressed frustration at the amount of data acquisition that is required for such "mature" studies. He noted that while some reviewer suggestions improve the work, "many suggested experiments [that] are unnecessary, and sometimes the requested work is so extensive that it constitutes a separate study onto itself."<sup>32</sup>

Vale's article preprint posted to bioRxiv.org resonated well within the scientific community by garnering thousands of views and hundreds of social media hits, and it was later published with peer review.<sup>33</sup> Vale's ultimate argument was for faster publication, particularly through publication of smaller studies and use of preprint servers. These solutions are consistent with the open science values of the data curation community. Pragmatically, more publication of "partial" studies would also likely yield smaller, more readily curated data sets; quicker time to data sharing could likely curb some information entropy. Nonetheless, the potential synergy could be wasted unless there is an effort to understand that the resistance to greater data sharing may have a deeper-seated resistance that is rooted in the broader data-related demands placed on researchers during other



parts of the research process. Our message has to be laser-focused on the value of data set availability and curation, and not focused simply on “data sharing” since so many within the scientific communities view article publication as data sharing already. Without addressing this critical nuance we may become just another voice seeming to arbitrarily demand that more time and effort be spent on data.

## *Overlooked Data Sharing: Supplemental Material*

Similarly, a form of data sharing often overlooked in the data curation community is supplemental material provided along with a published journal article, also known as supplemental data, auxiliary information, supporting information, or supplementary content.<sup>34</sup> Supplemental material is generally supplied to a publication in free form as an extra file or files that help support the main article. A prototypical example is a PDF that may include additional text, methods, analyses, figures, tables, and/or data, but other supplement examples may include file formats incompatible with article format or layout, such as video or code.<sup>35</sup> *PLOS* and *Science*, as just two examples, allow a myriad of file formats as supplemental files.

There are several reasons for providing supplemental material, such as allowing a reader to focus on the most salient points in the main body of the article or allowing the reader to access material that logistically cannot be placed within the main body due to size or format. Authors may submit supplements as a way to demonstrate that their work is thorough and well-executed, or they may submit under the belief that extra material may help “immunize” them from reviewer concerns.<sup>36</sup> Material that may have belonged in the main body is sometimes otherwise relegated to supplemental files due to journal space considerations or to minimize author page costs.<sup>37</sup>

Supplemental material is often tied to the advent of the electronic journal, but scientists have been providing more detail for primary articles via supplements for decades (for early examples in print see Myers and Abeles in 1990 and Sapp, Lord, and Hammarlund in 1975).<sup>38</sup> However, rapid adoption of electronic supplemental materials began in the late 1990s.<sup>39</sup> Over the course of a decade, Beauchamp of *The Journal of Clinical Investigation* reported that the percent of articles containing a supplement jumped from just 3 percent in 2001 to 95 percent in 2011.<sup>40</sup> Similar results have been reported for *The Journal of Experimental Medicine* and *The Journal of Neuroscience*.<sup>41</sup> Kenyon and Sprague’s thorough analysis of sixty journals broadly covering the environmental sciences similarly found that supplemental file adoption picked up quickly, albeit not entirely uniformly, between 2000 and 2011.<sup>42</sup>

The rapid adoption of supplemental materials suggests a successful data sharing mechanism; however, the practice has not been without debate. Libraries are concerned with the apparent “Pandora’s box of management issues” including a lack of document structure, metadata, persistence, and discoverability.<sup>43</sup> Journal editors have also weighed in with their concerns about quality, overhead, and the relevance of supplemental materials.<sup>44</sup> At least one journal has banned supplemental materials altogether over these concerns,<sup>45</sup> and others have implemented policies that limit supplemental materials to only that which is “essential.” On the other hand, many journals encourage supplements and also recommend a variety of file formats beyond just PDFs (e.g., see table 2 in Kenyon and Sprague).<sup>46</sup> In light of these inconsistencies and concerns, NISO and NFAIS established a formal working group in 2010 to develop recommended practices.<sup>47</sup> This group uncovered a messy landscape both in opinion and practice. Not only is the content highly variable, the handling of supplemental material by journals is idiosyncratic as well. For example, sometimes supplements are peer-reviewed, sometimes not; sometimes supplements and articles are formally linked, sometimes not. Culturally, Swartzman found two distinct camps: those who encouraged as much additional detail as deemed necessary, and those who felt supplemental materials were being used as a “data dump.”<sup>48</sup> Although one might be inclined to dismiss the concerns of journal editors as business-motivated rather than scientific-value-motivated, this is not the only arena where “overflow” concerns have emerged. During interviews with biomedical researchers, Siebert, Machesky, and Insall found that interviewees expressed many overflow-related concerns, including the proliferation of new journals, the explosion of publications, and even an excess of scientists themselves. This cumulated in an overarching concern that “rapid proliferation of scientific outputs was inconsistent with the capacity of the world of science to verify the quality of outputs.”<sup>49</sup>

Regardless of the greater scientific community’s ability to process the deluge of information, it’s clear that many scientists *are* willing to share additional information via supplements, and at least some portion of the scientific community appreciate the added content. Although supplemental materials may contain more than data, data curators’ skills squarely align with addressing the flaws of supplemental materials: unstructured information, lack of metadata, uncertain access persistence, and limited discoverability. Indeed, “Most frequently, supplemental materials suffer from a lack of descriptive metadata.”<sup>50</sup> As data curators we can view supplemental materials as a positive model and can pitch curation services as being able to alleviate several of the drawbacks that vex research communities. For communities that have embraced supplemental materials, one model may be able to encourage researchers to think of deposit into data repositories as “Upgraded Supplemental Materials,” where upgrade may mean something along a continuum of minimal metadata at one end to detailed curation at the other, depending on the scope of services available. Here we can emphasize consistent

metadata, persistent identifiers, stability, availability, and file format flexibility as directly addressing the nearly universally acknowledged limitations of supplemental files. While not perfect, even the most minimal, unmediated deposit is a huge step towards progress when compared to the current haphazard landscape of supplemental materials as described here.

## *Overrated Data Sharing: Unsustained Community Resources*

In juxtaposition to the unstructured nature of supplemental materials or the limitations of published articles, an untold number of highly structured and sophisticated data resources have also been developed. When the topic of domain repositories is broached, successful well-known examples such as Inter-university Consortium for Political and Social Research (ICPSR), GenBank, or the Sloan Digital Sky Survey quickly spring to mind; however, as of October 2015 the Registry of Research Data Repositories, re3data.org, contained 1,363 reviewed repositories with representation across both the humanities and sciences.<sup>51</sup> The number of resources represented in re3data.org is steadily growing, and it's understood that the registry is not yet comprehensive. For example, since 1993 *Nucleic Acids Research* has published an annual "Database Issue" and maintained an online Molecular Biology Database Collection that currently references 1,549 databases dedicated solely to bioinformatics and molecular biology.<sup>52</sup> Thus it's difficult to estimate how many data resources are currently available, but clearly data resources are of keen interest to many research communities.

Sustaining resources, however, is a much different animal. Established repositories are often asked to absorb endangered data, as recently occurred when the Cultural Policy and the Arts National Data Archive (CPANDA) began migration of data to the ICPSR and the National Archive of Data on Arts and Culture (NADAC) after conclusion of funding.<sup>53</sup> However, a lack of committed funding is a major concern for even the most successful and well-used domain repositories.<sup>54</sup> Recalcitrant funding agencies are extremely hesitant to commit to funding anything in perpetuity, citing their missions to spur innovation and the need to be responsive to new scientific directions. To be fair, the agencies are in a difficult position. As the number of new resources increases over time, the amount of funding required to sustain those resources would likewise accumulate. Without triage or alternative support mechanisms, undoubtedly funders fear that sustaining infrastructure will disproportionately result in reduced funding for new research.

This has created a habitual scenario where resources are left in limbo to scramble for support. In some cases, resources have been "sunsetting" due to lack

of community use or buy-in. Interestingly, in their exploration of data sharing behavior in the social sciences, Kim and Adler found that just because a data repository exists does not mean a community finds value in it.<sup>55</sup> One high profile example in the biological sciences is the Knowledgebase for the Protein Structure Initiative (PSI), a fifteen-year program funded through NIH's National Institute of General Medical Sciences, which aimed to advance technologies for the determination of three-dimensional protein structures. On conclusion of the PSI project, three review committees jointly concluded that the resource had yet to demonstrate broad use across the user communities and fate of the PSI Knowledgebase lies unknown.<sup>56</sup> Similar concerns were expressed for the recently sunsetted Virtual Astronomy Observatory.<sup>57</sup> When promises are made that such resources will empower scientific communities by providing access to data and yet the resources fail to live up to that promise, it's disillusioning to scientists who are already frustrated by the hypercompetitive funding climate. The arguments against these resources were that the money could be better spent elsewhere. Institutions with funds devoted to data curation and repositories meant to support data sharing are no less susceptible to such budgetary criticism at the local level; thus buy-in from local scientific communities is essential.

However, it's not only a lack of community buy-in that has doomed some resources. For example, in 2007 the National Library of Medicine announced plans to cut funding to five community resources and redirect funds towards "research and training." The resources had several thousand users, and communities attempted to rally in order to save them.<sup>58</sup> Likewise, the extremely popular Kyoto Encyclopedia of Genes and Genomes (KEGG) issued pleas in 2011 reminiscent of a National Public Radio pledge drive after restructuring of the primary Japanese funder.<sup>59</sup> Users who benefited from KEGG were urged "to write, email, tweet, and blog about your support for KEGG. I hope, in the long run, your voices will increase our chances of getting more stable funding."<sup>60</sup> KEGG has turned to a partly commercial model, but is still not fully sustainable. Time and again, resources have been put in peril despite demonstrated value to communities.

Notwithstanding the clear inability to sustain each new resource developed, researchers have had a penchant for developing such resources, frequently as a by-product of a larger research project (such as the PSI Knowledgebase as part of the larger PSI program described above). Likewise, funding agencies have a penchant for enabling such efforts, if not outright encouraging or requiring them. On one hand, these resources stand as further testaments to active data sharing. On the other hand, post-grant support planning has not been emphasized until recently, as evidenced by adoption of data management plans by federal grant agencies, and even today there has been no dissuasion of standing up isolated resources that will ultimately need migration, rescue, or sunseting. This has created a culture of at-risk data with no end in sight. These high-profile failures—

whether they represent lack of community use or lack of sustained funding—are another reason why scientists may be doubtful since they cast a shadow of hopelessness on data sharing. Indeed, such efforts begin to look overrated. One critical thing we can do as data curators is attempt to circumvent diversion of funds into one-off resources and instead emphasize the importance of centralized, community-based solutions whether they include our local institutional repositories or domain data repositories housed at other institutions. A major hurdle for us to achieve this will be aligning idiosyncratic needs of unique projects with the broad service models of community resources. Here, we can remind researchers that giving up the customization and control of a uniquely developed resource allows for more project funds and energy to go to the research at hand.

## *Overrated Data Sharing: Hyperbolic Arguments*

It is not a foregone conclusion that all data, even that without restrictions, should be shared. All data is not equally valuable, and several public access implementation plans have made it clear that they do not expect all data to be available. For example, the NIH states, “It is important to note that not all digital scientific data need to be shared and preserved.”<sup>61</sup> Likewise, NSF plan stated, “rarely does NSF expect that retention of all data that are streamed from an instrument or created in the course of an experiment or survey will be required.”<sup>62</sup> In fact, the OSTP memo itself expects that agency plans will take into account “preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden.”<sup>63</sup>

Not only is the data not always required to validate or reproduce research results, but the reuse utility varies dramatically between discipline, purpose of original study, and data types (e.g., see Borgman’s 2012 discussion of data types categorized as observational, computational, experimental, and records).<sup>64</sup> There is no universal approach, and broad data availability is not yet mature enough for ready identification of data that has enduring value. Furthermore, as Borgman noted, “Perhaps the most significant challenge to data sharing is the lack of demonstrated demand for research data outside of genomics, climate science, astronomy, social science surveys, and a few other areas.”<sup>65</sup> This is a reality that dramatically complicates the data-sharing landscape. Efforts such as the Stewardship Gap Project aim to clarify this reality by developing evaluation frameworks and recommendations to identify data of particularly high value along with the support required to ensure long-term access.<sup>66</sup> Because of the current ambiguity, however, overemphasis on the impact of data specifically may also confuse or even aggravate some researchers.

Rather than reuse the data, some data will simply be replaced because the original data has only transitory use for a specific experiment. Take an example of observing the growth of a bacteria population. One way to measure growth is to inoculate a liquid culture with a very small amount of “starter” bacteria from a pure stock of bacteria. The liquid starts out clear, and the researcher essentially measures the increase in “cloudiness” of the liquid as the bacteria grow over time. The raw data is a series of time points and the density (“cloudiness”) measurement at each of those times, which is then represented as a graph of density ( $y$ -axis) versus time ( $x$ -axis). If the wrong bacterial stock was mistakenly used to conduct the growth experiment, neither the raw data nor the graphical representation necessarily divulge that error since it’s just a measure of bacterial density and not of bacterial type. While for some types of research, access to raw data in its original format may be helpful or even imperative, this is an example where the underlying data likely holds no more utility than representations of the data. Accounting for error and fluctuation is why independent replication *within a given study* is critical and considered a cornerstone in experimental sciences.<sup>67</sup> Should other researchers want to replicate the initial findings, they would never reuse the raw data by replotting the graph of growth. They would redo the entire experiment and acquire their own growth measurements independently to account for potential flaws or idiosyncrasies in the researcher’s execution, protocol, materials, or environment. It’s not a matter of trust in the data; it’s a matter of external verification of the experiment as a whole. In fact, Crotty and commenters argue that clear and accurate methodology is more important than data access.<sup>68</sup> On the other hand, the very same project may include a genomic analysis of the bacterial culture, and the ensuing genomic sequences may be of reuse utility. Unfortunately, because no absolutes apply, we simply cannot state that data sharing practices *are* appropriate for one data type or *are not* appropriate for another data type, even within a given discipline. It is maddeningly messy.

While scientific communities, agencies, and publishers struggle to establish which data to share or not share, scientists may feel obligated to share everything, regardless of value, which evokes the “data dump” concern already associated with supplemental materials. While perhaps overcompensation is an enviable problem, the issue of long-term value will be further exacerbated by the continued lack of definitions, standards, and best practices, which are all equally important but even more difficult to address. If some scientists share not because they—or anyone else—*truly value the data* but simply because they view data sharing as insulation against criticism or as a requirement for compliance, we have to prepare ourselves in the data curation community to ask: does this data also warrant the substantial effort of curation and preservation? We must view scientists, both as consumers and producers of data, as our best partners in determining which data should benefit from our resources and for how long.

Because the data itself is just one component of research, a single-minded focus on data can ultimately detract from increased transparency and reproducibility. Without robust experimental design, such as use of proper controls and sampling procedures, raw data may be just as erroneous as a representative figure. Likewise, simulation data is critically dependent on the software versions used, the initial parameters used in a simulation run, and the general operating variables. If data sharing alone were to become a sort of rubber stamp for better research, large swaths of science will fail in this assessment. For these reasons, all disciplines have not necessarily taken the same path towards data sharing. In 2015, NIH issued plans to enhance rigor and transparency through four major areas: (1) the scientific premise of the proposed research, (2) rigorous experimental design for robust and unbiased results, (3) consideration of relevant biological variables, and (4) authentication of key biological and/or chemical resources.<sup>69</sup> Although NIH acknowledges that data is important, clearly it is not an all-encompassing solution. In this regard when the ultimate goal is to enable better science, then the best scenario is to enable inclusion of whatever has been missing, whether that be data, code, methodology, materials, or any other information. While in some cases the term *data* has become a bucket for anything research-related that's not a journal article, acknowledging semantic differences is important for the sake of productive communication and grittier issues like the application of intellectual property law. As mentioned above in the supplemental materials section, data in the "factual material" sense is not the only thing that could benefit from best practices, standardization, and curation. While this could be a potential complication to data curation services, data curators do not necessarily have to play an active role in hands-on curation of all things research-related, especially in the short term. Simply being knowledgeable of current and emerging trends, such as new policies and new sharing platforms, is of value. Indeed, such a role aligns with the reference services that stand as a fundamental mission of libraries. The benefit of thinking more broadly will be useful in the long term to the data curation profession, however, because accumulated knowledge through such conversations will enable user-informed evolution of data curation service models.

## Conclusions

While the data curation community has been justifiably buoyed by the impact of data sharing success stories, the points presented are intended to serve as examples of the nuances that surround data sharing. As data curators, we do ourselves a disservice if we look at data sharing only from the perspective of progressive or idealist attitudes. Without attempting to understand and accommodate the nuances of data sharing, then the lack of rapid, dedicated, and widespread adoption



of new practices will lead to frustration in the data curation community. Indeed, some antagonistic views, such as accusing scientists of misconduct, laziness, or lack of creativity if they fail to see a need for data sharing, have already surfaced in the back channels of the data curation community (e.g., social media, Listservs, and conferences),\* which may be a manifestation of frustration. Instead of setting ourselves up for disappointment, a more nimble approach is to acknowledge a broader perspective that stems from the variability of definitions, communities, practices, and science itself. For those who interface directly with scientists, ultimately our greatest effectiveness will come by virtue of working within the realities that scientists experience.

For example, the author received an e-mail some months ago from a faculty member who inquired if university-wide data sharing practices had been established. A publisher was requesting that individual-level data be made available, but the faculty member was reluctant to share. In the e-mail, the researcher initially cited the need to do a secondary analysis, the limitations of the data set, and the desire to share the data within the specific research community (as opposed to untargeted sharing) as reasons for not wanting to share openly. At first pass, some data sharing advocates would not find any of these reasons “valid.” A colleague and I met with the faculty member and two graduate students also on the project, and we devoted our time to simply listening and learning about their concerns. We learned that the publisher’s data sharing policies had changed mid-peer-review, and the faculty member held deep reservations about whether publishers, who may not be as attuned to data utility or as thoughtful of sharing consequences, are appropriate drivers of data sharing practices. We also learned that human subject participants had signed consents that stated data would be shared only in aggregate, which would mean time-consuming and potentially impossible re-consent of each participant prior to sharing deidentified participant-level data. Furthermore, if data was published from the study, the lack of accompanying control data would dramatically reduce utility. Perhaps most interestingly, we also learned that this research area had already established a committee to define best practices for data analysis and sharing, in which the faculty member participated, and a recommendations report was currently under community review. In truth, we found that faculty member was a supporter of data sharing, but felt strongly that sharing at all costs was senseless. Indeed, it

---

\* For example, at the 2016 International Digital Curation Conference, a keynote address described supplemental files as “malpractice” (Barend Mons, “Open Science as a Social Machine: Where (the...) Are the Data?” [keynote address, International Digital Curation Conference, Amsterdam, the Netherlands, February 22–25, 2016], <http://www.dcc.ac.uk/sites/default/files/documents/IDCC16/Keynotes/Barend%20Mons.pdf>), and “data whining” emerged on Twitter during one panel, for example “Lots of talk at this #IDCC16 Panel session on data whining (instead of data mining). All the reasons why people can’t share their data...” (from #IDCC16 hashtag archive at <http://bit.ly/1RsVJzt> via @alastairdunning).



was also our conclusion that the cost-benefit ratio of sharing in this case was unfavorable, and we recommended the faculty member request an exception from the editor, which ultimately proved successful. The data was not shared. Had we taken the view that unwavering promotion of data sharing is the only acceptable position, it's likely that we would have failed in establishing ourselves as credible resource. Instead, we gained the faculty member's confidence as balanced and knowledgeable professionals who are supportive of research as a whole. Notably, through our interactions the group has now adopted language for participant consent that will allow for more facile and permissive data sharing in the future.

While we must keep in mind that current practices are not uniformly contested, nor is data sharing a universal panacea, it is clear that sharing will become more commonplace in coming years. There is no doubt that data curation has had—and will continue to have—an important place in science. As data sharing practices evolve, data curators have the opportunity to craft our message and services in a way that both makes sense and delivers great value to the communities we aim to serve. The strategies include (1) acknowledging cultural pressures and norms, (2) providing directness and clarity in messaging to emphasize purpose, (3) seeking to augment or enhance current practices, and (4) embracing and planning for complexity. While such strategies may fall short of ideals, they place data curators in a position to enable more efficient and robust science through closer alignment with research communities.

## Acknowledgments

The author would like to thank Elise Dunham, Bill Mischo, Sarah Williams, editor Lisa Johnston, and anonymous reviewers for thoughtful and critical evaluation of this chapter.

## Notes

1. Anna K. Gold, "Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries," *D-Lib Magazine* 13, no. 9/10 (2007), <http://works.bepress.com/agold01/4/>.
2. Christine L. Borgman, "The Conundrum of Sharing Research Data," *Journal of the American Society for Information Science and Technology* 63, no. 6 (June 2012): 1059–78, doi:10.1002/asi.22634.
3. Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame, "Data Sharing by Scientists: Practices and Perceptions," *PLoS ONE* 6, no. 6 (2011): e21101, doi:10.1371/journal.pone.0021101; Carol Tenopir, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett, "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide," *PLoS ONE*

- 10, no. 8 (2015): e0134826, doi:10.1371/journal.pone.0134826; Liz Ferguson, “How and Why Researchers Share Data (and Why They Don’t),” *Exchanges* (blog), November 3, 2014, <https://web.archive.org/web/20160116150325/http://exchanges.wiley.com/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont/>; Sarah C. Williams, “Data Sharing Interviews with Crop Sciences Faculty: Why They Share Data and How the Library Can Help,” *Issues in Science and Technology Librarianship*, Spring 2013, doi:10.5062/F4T151M8.
4. Robert S. Decker, Leslie Wimsatt, Andrea G. Trice, and Joseph A. Konstan, *A Profile of Federal-Grant Administrative Burden among Federal Demonstration Partnership Faculty*, “A Report of the Faculty Standing Committee of the Federal Demonstration Partnership (Federal Demonstration Partnership, January 2007), [http://web.archive.org/web/20160214195603/http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pgasite\\_054586.pdf](http://web.archive.org/web/20160214195603/http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pgasite_054586.pdf).
  5. Sandra L. Schneider, Kirsten K. Ness, Sara Rockwell, Kelly Shaver, and Randy Brutkiewicz, *2012 Faculty Workload Survey*, research report (Federal Demonstration Partnership, April 2014), [http://web.archive.org/web/20151022202705/http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pgasite\\_087667.pdf](http://web.archive.org/web/20151022202705/http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pgasite_087667.pdf).
  6. National Institutes of Health, “Success Rates and Funding Rates. Research Project Grants: Competing Applications, Awards, and Success Rates 1998–2014,” NIH Data Book, 2015, <http://web.archive.org/web/20151022204459/http://report.nih.gov/NIDatobook/Charts/Default.aspx?showm=Y&chartId=124&catId=13>; National Science Foundation, *Report to the National Science Board on the National Science Foundation’s Merit Review Process: Fiscal Year 2013* (Arlington, VA: National Science Foundation, May 2014), <http://web.archive.org/web/20151022211742/https://www.nsf.gov/nsb/publications/2014/nsb1432.pdf>.
  7. IPAMM Working Group, *Impact of Proposal and Award Management Mechanisms* (Arlington, VA: National Science Foundation, 2007), <http://web.archive.org/web/20151024203046/http://www.nsf.gov/pubs/2007/nsf0745/nsf0745.pdf>.
  8. National Institutes of Health, “Definitions of Criteria and Considerations for Research Project Grant (RPG/X01/R01/R03/R21/R33/R34) Critiques,” National Institutes of Health Grants and Funding, last updated March 21, 2016, [http://web.archive.org/web/20160325020441/https://grants.nih.gov/grants/peer/critiques/rpg\\_D.htm](http://web.archive.org/web/20160325020441/https://grants.nih.gov/grants/peer/critiques/rpg_D.htm).
  9. Tenopir et al., “Changes in Data Sharing.”
  10. Jeremy Berg, “Indirect Cost Rate Survey,” *Data Hound* (blog), May 10, 2014, <https://web.archive.org/web/20150924200621/http://datahound.scientopia.org/2014/05/10/indirect-cost-rate-survey/>.
  11. Burroughs Wellcome Fund, “Obtaining Tenure,” Career Tools, 2014, <http://web.archive.org/web/20151024181949/http://www.bwfund.org/career-tools/obtaining-tenure>; Chelsea Wald, “Redefining Tenure at Medical Schools,” *Science Careers*, March 6, 2009, doi:10.1126/science.caredit.a0900032.
  12. John P. Holdren, “Increasing Access to the Results of Federally Funded Scientific Research,” Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013, [http://web.archive.org/web/20160115125401/https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://web.archive.org/web/20160115125401/https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
  13. US Geological Survey, “Scientific Data Management Foundation, US Geological Survey Instructional Memorandum No. IM OSQI 2015-01, US Geological Survey Manual,

- February 19, 2015, <http://web.archive.org/web/20151031000659/http://www.usgs.gov/usgs-manual/im/IM-OSQI-2015-01.html>.
14. C. Titus Brown, "Cultural Confusions about Data: The Intertidal Zone between Two Styles of Biology," *Living in an Ivory Basement* (blog), April 2, 2015, <http://web.archive.org/web/20151003163047/http://ivory.idyll.org/blog/2015-culturally-confused-about-data.html>.
  15. Holdren, "Increasing Access to the Results of Federally Funded Scientific Research."
  16. Allen H. Renear, Simone Sacchi, and Karen M. Wickett, "Definitions of *Dataset* in the Scientific and Technical Literature," *Proceedings of the American Society for Information Science and Technology* 47, no. 1 (2010): 1–4, doi:10.1002/meet.14504701240.
  17. Tenopir et al., "Data Sharing by Scientists."
  18. Tenopir et al., "Changes in Data Sharing."
  19. Ferguson, "How and Why Researchers Share Data."
  20. Holdren, "Increasing Access to the Results of Federally Funded Scientific Research."
  21. Caroline J. Savage and Andrew J. Vickers, "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals," *PLoS ONE* 4, no. 9 (2009): e7078, doi:10.1371/journal.pone.0007078; Jelte M. Wicherts, Denny Borsboom, Judith Kats, and Dylan Moleenaar, "The Poor Availability of Psychological Research Data for Reanalysis," *American Psychologist* 61, no. 7 (October 2006): 726–28, doi:10.1037/0003-066X.61.7.726.
  22. Williams, "Data Sharing Interviews with Crop Sciences Faculty"; Philip Herold, "Data Sharing among Ecology, Evolution, and Natural Resources Scientists: An Analysis of Selected Publications," *Journal of Librarianship and Scholarly Communication* 3, no. 2 (2015): eP1244, doi:10.7710/2162-3309.1244.
  23. Ajai R. Singh and Shakuntala A. Singh, "Ethical Obligation towards Research Subjects," *Mens Sana Monographs* 5, no. 1 (2007): 107–12, doi:10.4103/0973-1229.32153.
  24. Sylvia M. Burwell, Steven VanRoekel, Todd Park, and Dominic J. Mancini, "Open Data Policy: Managing Information as an Asset," Memorandum for the Heads of Executive Departments and Agencies, Office of Management and Budget, May 9, 2013, <https://web.archive.org/web/20151104005245/https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>.
  25. Inter-university Consortium for Political and Social Research (ICPSR), "Phase 5: Preparing Data for Sharing," in *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle*, 5th ed. (Ann Arbor, MI: ICPSR, 2012), 5.
  26. National Institutes of Health. *Data Sharing Workbook* (Bethesda, MD: National Institutes of Health, last revised February 13, 2004), [http://web.archive.org/web/20151116180118/https://grants.nih.gov/grants/policy/data\\_sharing/data\\_sharing\\_workbook.pdf](http://web.archive.org/web/20151116180118/https://grants.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf).
  27. National Science Foundation, *Data Management for NSF EHR Directorate* (Arlington, VA: National Science Foundation Education and Human Resources Directorate, March 2011), <http://web.archive.org/web/20151030234444/http://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf>; US Geological Survey, "Scientific Data Management Foundation."
  28. William Mischo, Mary Schlemback, and Megan O'Donnell, "An Analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals," *Journal of eScience Librarianship* 3, no. 1 (2014), doi:10.7191/jeslib.2014.1060.
  29. Youngseek Kim and Melissa Adler, "Social Scientists' Data Sharing Behaviors: Investigating the Roles of Individual Motivations, Institutional Pressures, and Data Repositories," *International Journal of Information Management* 35, no. 4 (August 2015): 408–18, doi:10.1016/j.ijinfomgt.2015.04.007.

30. Ronald D. Vale, "Accelerating Scientific Publication in Biology," preprint, *bioRxiv*, September 12, 2015, doi:10.1101/022368.
31. Sabina Siebert, Laura M. Machesky, and Robert H. Insall, "Overflow in Science and Its Implications for Trust," *eLife* 4 (2015): e10825, doi:10.7554/eLife.10825.
32. Vale, "Accelerating Scientific Publication in Biology."
33. Ronald D. Vale, "Accelerating Scientific Publication in Biology," *Proceedings of the National Academy of Sciences* 112 (2015): 13,439–46, doi:10.1073/pnas.1511912112.
34. Jeremy Kenyon and Nancy R. Sprague, "Trends in the Use of Supplementary Materials in Environmental Science Journals," *Issues in Science and Technology Librarianship*, Winter 2014, doi:10.5062/F40Z717Z.
35. Alexander Schwarzman, "Supplemental Information: Who's Doing What and Why" (PowerPoint presentation, CSE 2012 Annual Meeting, Seattle, WA, May 20, 2012), [http://web.archive.org/web/20151013195224/http://www.niso.org/apps/group\\_public/document.php?document\\_id=8591&wg\\_abbrev=supptechnical](http://web.archive.org/web/20151013195224/http://www.niso.org/apps/group_public/document.php?document_id=8591&wg_abbrev=supptechnical); Linda Beebe, "Supplemental Materials for Journal Articles: NISO/NFAIS Joint Working Group," *Information Standards Quarterly* 22, no. 3 (Summer 2010): 33–37, [http://web.archive.org/web/20151227001104/http://www.niso.org/apps/group\\_public/download.php/4885/Beebe\\_SuppMatls\\_WG\\_ISQ\\_v22no3.pdf](http://web.archive.org/web/20151227001104/http://www.niso.org/apps/group_public/download.php/4885/Beebe_SuppMatls_WG_ISQ_v22no3.pdf).
36. John Maunsell, "Announcement Regarding Supplemental Material," *Journal of Neuroscience* 30 (2010): 10,599–600.
37. American Physical Society, "Supplemental Data," *APS Online Style Manual*, 2003, <http://web.archive.org/web/20160325000016/http://www.apstylemanual.org/oldmanual/parts/supplemental.htm>.
38. Robert W. Myers and Robert H. Abeles, "Conversion of 5-S-Methyl-5-Thio-D-Ribose to Methionine in *Klebsiella Pneumoniae*. Stable Isotope Incorporation Studies of the Terminal Enzymatic Reactions in the Pathway," *Journal of Biological Chemistry* 265 (1990): 16,913–21; Curtis Sapp, Michael Lord, and E. Roy Hammarlund, "Sodium Chloride Equivalents, Cryoscopic Properties, and Hemolytic Effects of Certain Medicinals in Aqueous Solution III: Supplemental Values," *Journal of Pharmaceutical Sciences* 64, no. 11 (November 1975): 1884–86, doi:10.1002/jps.2600641132.
39. Nicholas R. Anderson, Peter Tarczy-Hornoch, and Roger E. Bumgarner, "On the Persistence of Supplementary Resources in Biomedical Publications," *BMC Bioinformatics* 7 (2006): 260, doi:10.1186/1471-2105-7-260; Maunsell, "Announcement Regarding Supplemental Material."
40. Alexander Schwarzman, "Supplemental Materials Survey," *Information Standards Quarterly* 22, no. 3 (Summer 2010): 23–26, [http://web.archive.org/web/20151013194617/http://www.niso.org/apps/group\\_public/download.php/4886/Schwarzman\\_SuppMatls-Survey\\_ISQ\\_v22no3.pdf](http://web.archive.org/web/20151013194617/http://www.niso.org/apps/group_public/download.php/4886/Schwarzman_SuppMatls-Survey_ISQ_v22no3.pdf).
41. Christine Borowski, "Enough Is Enough," *Journal of Experimental Medicine* 208, no. 7 (2011): 1337, doi:10.1084/jem.20111061; Maunsell, "Announcement Regarding Supplemental Material."
42. Kenyon and Sprague, "Trends in the Use of Supplementary Materials."
43. Todd Carpenter, "Supplementary Materials: A Pandora's Box of Issues Needing Best Practices," *Against the Grain* 21 (2009): 84–85; Thomas Schaffer and Kathy M. Jackson, "The Use of Online Supplementary Material in High-Impact Scientific Journals," *Science and Technology Libraries* 25, no. 1–2 (2004): 73–85, doi:10.1300/J122v25n01\_06.
44. Emilie Marcus, "Taming Supplemental Material," *Neuron* 64, no. 1 (October 2009):

- 3, doi:10.1016/j.neuron.2009.09.046; Borowski, “Enough Is Enough”; Maunsell, “Announcement Regarding Supplemental Material.”
45. Maunsell, “Announcement Regarding Supplemental Material.”
  46. Kenyon and Sprague, “Trends in the Use of Supplementary Materials.”
  47. Beebe, “Supplemental Materials for Journal Articles.”
  48. Schwarzman, “Supplemental Information.”
  49. Siebert, Machesky, and Insall, “Overflow in Science.”
  50. Beebe, “Supplemental Materials for Journal Articles.”
  51. Heinz Pampel, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, et al., “Making Research Data Repositories Visible: The re3data.org Registry,” *PLoS ONE* 8, no. 11 (2013): e78080, doi:10.1371/journal.pone.0078080.
  52. Michael Y. Galperin, Daniel J. Rigden, and Xosé M. Fernández-Suárez, “The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection,” *Nucleic Acids Research* 43, no. D1 (2015): D1–5, doi:10.1093/nar/gku1241.
  53. CPANDA, “What Is CPANDA?” 2015, <http://web.archive.org/web/20151120151725/http://www.cpanda.org/cpanda/about>.
  54. Carol Ember and Robert Hanisch, “Sustaining Domain Repositories for Digital Data: A White Paper,” 2013, [http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper\\_ICPSR\\_SDRDD\\_121113.pdf](http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf).
  55. Kim and Adler, “Social Scientists’ Data Sharing Behaviors.”
  56. Future of Structural Biology Committees, *Recommendations for Continued Investment in Structural Biology Following the Sunset of the Protein Structure Initiative* (Bethesda, MD: National Institute of General Medical Sciences, December 2014), <http://web.archive.org/web/20151013183712/http://www.nigms.nih.gov/News/reports/Documents/NIGMS-FSBC-report2014.pdf>.
  57. Stephen Kent, Chryssa Kouveliotou, David Meyer, Richard H. Miller, David Schade, James Schombert, Alexander Szalay, and Suresh Santhana Vannan, *Report of the Astrophysics Archives Program Review for the Astrophysics Division, Science Mission Directorate* (NASA, May 6–8, 2015), <http://web.archive.org/web/20151107185033/http://science.nasa.gov/media/medialibrary/2015/07/08/NASA-AAPR2015-FINAL.pdf>.
  58. Monya Baker, “Databases Fight Funding Cuts,” *Nature* 489 (2012): 19, doi:10.1038/489019a.
  59. Minoru Kanehisa, “Plea to Support KEGG,” *Kyoto Encyclopedia of Genes and Genomes*, 2011, <http://web.archive.org/web/20151108212102/http://www.genome.jp/kegg/docs/plea.html>.
  60. *Ibid.*
  61. National Institutes of Health, *Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research* (Bethesda, MD: National Institutes of Health, February 2015), <https://web.archive.org/web/20150908072046/https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>.
  62. National Science Foundation, *Today’s Data, Tomorrow’s Discoveries*, NSF 15-52 (Arlington, VA: National Science Foundation, 2015), <https://web.archive.org/web/20160131120745/http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf>.
  63. Holdren, “Increasing Access to the Results of Federally Funded Scientific Research.”
  64. Borgman, “The Conundrum of Sharing Research Data.”
  65. *Ibid.*

66. Stewardship Gap Project, “The Problem,” 2015, [http://web.archive.org/web/20160214022939/http://www.colorado.edu/ibs/cupc/stewardship\\_gap/](http://web.archive.org/web/20160214022939/http://www.colorado.edu/ibs/cupc/stewardship_gap/).
67. David L. Vaux, “Research Methods: Know When Your Numbers Are Significant,” *Nature* 492 (2012): 180–81, doi:10.1038/492180a.
68. David Crotty, “Nevermind the Data, Where Are the Protocols?” *The Scholarly Kitchen* (blog), November 18, 2014, <https://web.archive.org/web/20160201153044/http://scholarlykitchen.sspnet.org/2014/11/18/nevermind-the-data-where-are-the-protocols/>.
69. National Institutes of Health, “Enhancing Reproducibility through Rigor and Transparency,” NOT-OD-15-103, October 30, 2015, <http://web.archive.org/web/20151030024011/http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html>.

## Bibliography

- American Physical Society. “Supplemental Data.” *APS Online Style Manual*, 2003. <http://web.archive.org/web/20160325000016/http://www.apstylemanual.org/oldmanual/parts/supplemental.htm>.
- Anderson, Nicholas R., Peter Tarczy-Hornoch, and Roger E. Bumgarner. “On the Persistence of Supplementary Resources in Biomedical Publications.” *BMC Bioinformatics* 7 (2006): 260. doi:10.1186/1471-2105-7-260.
- Baker, Monya. “Databases Fight Funding Cuts.” *Nature* 489 (2012): 19. doi:10.1038/489019a.
- Beebe, Linda. “Supplemental Materials for Journal Articles: NISO/NFAIS Joint Working Group.” *Information Standards Quarterly* 22, no. 3 (Summer 2010): 33–37. [http://web.archive.org/web/20151227001104/http://www.niso.org/apps/group\\_public/download.php/4885/Beebe\\_SuppMatls\\_WG\\_ISQ\\_v22no3.pdf](http://web.archive.org/web/20151227001104/http://www.niso.org/apps/group_public/download.php/4885/Beebe_SuppMatls_WG_ISQ_v22no3.pdf).
- Berg, Jeremy. “Indirect Cost Rate Survey.” *Data Hound* (blog), May 10, 2014. <https://web.archive.org/web/20150924200621/http://datahound.scientopia.org/2014/05/10/indirect-cost-rate-survey/>.
- Borgman, Christine L. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63, no. 6 (June 2012): 1059–78. doi:10.1002/asi.22634.
- Borowski, Christine. “Enough Is Enough.” *Journal of Experimental Medicine* 208, no. 7 (2011): 1337. doi:10.1084/jem.20111061.
- Brown, C. Titus. “Cultural Confusions about Data: The Intertidal Zone between Two Styles of Biology.” *Living in an Ivory Basement* (blog), April 2, 2015. <http://web.archive.org/web/20151003163047/http://ivory.idyll.org/blog/2015-culturally-confused-about-data.html>.
- Burroughs Wellcome Fund. “Obtaining Tenure.” Career Tools, 2014. <http://web.archive.org/web/20151024181949/http://www.bwfund.org/career-tools/obtaining-tenure>.
- Burwell, Sylvia M., Steven VanRoekel, Todd Park, and Dominic J. Mancini. “Open Data Policy: Managing Information as an Asset.” Memorandum for the Heads of Executive Departments and Agencies, Office of Management and Budget, May 9, 2013. <https://web.archive.org/web/20151104005245/https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>.



- Carpenter, Todd. "Supplementary Materials: A Pandora's Box of Issues Needing Best Practices." *Against the Grain* 21 (2009): 84–85.
- CPANDA. "What Is CPANDA?" 2015. <http://web.archive.org/web/20151120151725/http://www.cpanda.org/cpanda/about>.
- Crotty, David. "Nevermind the Data, Where Are the Protocols?" *The Scholarly Kitchen* (blog), November 18, 2014. <https://web.archive.org/web/20160201153044/http://scholarly-kitchen.sspnet.org/2014/11/18/nevermind-the-data-where-are-the-protocols/>.
- Decker, Robert S., Leslie Wimsatt, Andrea G. Trice, and Joseph A. Konstan. *A Profile of Federal-Grant Administrative Burden among Federal Demonstration Partnership Faculty*. A Report of the Faculty Standing Committee of the Federal Demonstration Partnership. Federal Demonstration Partnership, January 2007. [http://web.archive.org/web/20160214195603/http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pgasite\\_054586.pdf](http://web.archive.org/web/20160214195603/http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pgasite_054586.pdf).
- Ember, Carol, and Robert Hanisch. "Sustaining Domain Repositories for Digital Data: A White Paper," 2013. [http://datacommunity.icpsr.umich.edu/sites/default/files/White-Paper\\_ICPSR\\_SDRDD\\_121113.pdf](http://datacommunity.icpsr.umich.edu/sites/default/files/White-Paper_ICPSR_SDRDD_121113.pdf).
- Ferguson, Liz. "How and Why Researchers Share Data (and Why They Don't)." *Exchanges* (blog), November 3, 2014. <https://web.archive.org/web/20160116150325/http://exchanges.wiley.com/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont/>.
- Future of Structural Biology Committees. *Recommendations for Continued Investment in Structural Biology Following the Sunset of the Protein Structure Initiative*. Bethesda, MD: National Institute of General Medical Sciences, December 2014. <http://web.archive.org/web/20151013183712/http://www.nigms.nih.gov/News/reports/Documents/NIGMS-FSBC-report2014.pdf>.
- Galperin, Michael Y., Daniel J. Rigden, and Xosé M. Fernández-Suárez. "The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection." *Nucleic Acids Research* 43, no. D1 (2015): D1–5. doi:10.1093/nar/gku1241.
- Gold, Anna K. "Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries." *D-Lib Magazine* 13, no. 9/10 (2007). <http://works.bepress.com/agold01/4/>.
- Herold, Philip. "Data Sharing among Ecology, Evolution, and Natural Resources Scientists: An Analysis of Selected Publications." *Journal of Librarianship and Scholarly Communication* 3, no. 2 (2015): eP1244. doi:10.7710/2162-3309.1244.
- Holdren, John P. "Increasing Access to the Results of Federally Funded Scientific Research." Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013. [http://web.archive.org/web/20160115125401/https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://web.archive.org/web/20160115125401/https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
- Inter-university Consortium for Political and Social Research (ICPSR). "Phase 5: Preparing Data for Sharing." In *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle*, 5th ed, 36–39. Ann Arbor, MI: ICPSR, 2012.
- IPAMM Working Group. *Impact of Proposal and Award Management Mechanisms: Final Report*. Arlington, VA: National Science Foundation, 2007. <http://web.archive.org/web/20151024203046/http://www.nsf.gov/pubs/2007/nsf0745/nsf0745.pdf>.

- Kanehisa, Minoru. "Plea to Support KEGG." *Kyoto Encyclopedia of Genes and Genomes*, May 16, 2011. <http://web.archive.org/web/20151108212102/http://www.genome.jp/kegg/docs/plea.html>.
- Kent, Stephen, Chryssa Kouveliotou, David Meyer, Richard H. Miller, David Schade, James Schombert, Alexander Szalay, and Suresh Santhana Vannan. *Report of the Astrophysics Archives Program Review for the Astrophysics Division, Science Mission Directorate*. NASA, May 6–8, 2015. <http://web.archive.org/web/20151107185033/http://science.nasa.gov/media/medialibrary/2015/07/08/NASA-AAPR2015-FINAL.pdf>.
- Kenyon, Jeremy, and Nancy R. Sprague. "Trends in the Use of Supplementary Materials in Environmental Science Journals." *Issues in Science and Technology Librarianship*, Winter 2014. doi:10.5062/F40Z717Z.
- Kim, Youngseek, and Melissa Adler. "Social Scientists' Data Sharing Behaviors: Investigating the Roles of Individual Motivations, Institutional Pressures, and Data Repositories." *International Journal of Information Management* 35, no. 4 (August 2015): 408–18. doi:10.1016/j.ijinfomgt.2015.04.007.
- Marcus, Emilie. "Taming Supplemental Material." *Neuron* 64, no. 1 (October 2009): 3. doi:10.1016/j.neuron.2009.09.046.
- Maunsell, John. "Announcement Regarding Supplemental Material." *Journal of Neuroscience* 30 (2010): 10,599–600.
- Mischo, William, Mary Schlembach, and Megan O'Donnell. "An Analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals." *Journal of eScience Librarianship* 3, no. 1 (2014). doi:10.7191/jeslib.2014.1060.
- Mons, Barend. "Open Science as a Social Machine: Where (the...) Are the Data?" [keynote address, International Digital Curation Conference, Amsterdam, the Netherlands, February 22–25, 2016], <http://www.dcc.ac.uk/sites/default/files/documents/IDCC16/Keynotes/Barend%20Mons.pdf>.
- Myers, Robert W., and Robert H. Abeles. "Conversion of 5-S-Methyl-5-Thio-D-Ribose to Methionine in *Klebsiella Pneumoniae*. Stable Isotope Incorporation Studies of the Terminal Enzymatic Reactions in the Pathway." *Journal of Biological Chemistry* 265 (1990): 16,913–21.
- National Institutes of Health. *Data Sharing Workbook*. Bethesda, MD: National Institutes of Health, last revised February 13, 2004. [http://web.archive.org/web/20151116180118/https://grants.nih.gov/grants/policy/data\\_sharing/data\\_sharing\\_workbook.pdf](http://web.archive.org/web/20151116180118/https://grants.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf).
- . "Definitions of Criteria and Considerations for Research Project Grant (RPG/X01/R01/R03/R21/R33/R34) Critiques." National Institutes of Health Grants and Funding, last updated March 21, 2016. [http://web.archive.org/web/20160325020441/https://grants.nih.gov/grants/peer/critiques/rpg\\_D.htm](http://web.archive.org/web/20160325020441/https://grants.nih.gov/grants/peer/critiques/rpg_D.htm).
- . "Enhancing Reproducibility through Rigor and Transparency," NOT-OD-15-103. October 30, 2015. <http://web.archive.org/web/20151030024011/http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-103.html>.
- . *Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research*. Bethesda, MD: National Institutes of Health, February 2015. <https://web.archive.org/web/20150908072046/https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>.
- . "Success Rates and Funding Rates: Research Project Grants: Competing Applications, Awards, and Success Rates 1998–2014." NIH Data Book, 2015. <http://web>.



- archive.org/web/20151022204459/http://report.nih.gov/NIHDataBook/Charts/Default.aspx?showm=Y&chartId=124&catId=13.
- National Science Foundation. *Data Management for NSF EHR Directorate: Proposals and Awards*. Arlington, VA: National Science Foundation Education and Human Resources Directorate, March 2011. <http://web.archive.org/web/20151030234444/http://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf>.
- . *Report to the National Science Board on the National Science Foundation's Merit Review Process: Fiscal Year 2013*. Arlington, VA: National Science Foundation, May 2014. <http://web.archive.org/web/20151022211742/https://www.nsf.gov/nsb/publications/2014/nsb1432.pdf>.
- . *Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation*. NSF 15-52. Arlington, VA: National Science Foundation, 2015. <https://web.archive.org/web/20160131120745/http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf>.
- Pampel, Heinz, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf. "Making Research Data Repositories Visible: The re3data.org Registry." *PLoS ONE* 8, no. 11 (2013): e78080. doi:10.1371/journal.pone.0078080.
- Renear, Allen H., Simone Sacchi, and Karen M. Wickett. "Definitions of *Dataset* in the Scientific and Technical Literature." *Proceedings of the American Society for Information Science and Technology* 47, no. 1 (2010): 1–4. doi:10.1002/meet.14504701240.
- Sapp, Curtis, Michael Lord, and E. Roy Hammarlund. "Sodium Chloride Equivalents, Cryoscopic Properties, and Hemolytic Effects of Certain Medicinals in Aqueous Solution III: Supplemental Values." *Journal of Pharmaceutical Sciences* 64, no. 11 (November 1975): 1884–86. doi:10.1002/jps.2600641132.
- Savage, Caroline J., and Andrew J. Vickers. "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals." *PLoS ONE* 4, no. 9 (2009): e7078. doi:10.1371/journal.pone.0007078.
- Schaffer, Thomas, and Kathy M. Jackson. "The Use of Online Supplementary Material in High-Impact Scientific Journals." *Science and Technology Libraries* 25, no. 1–2 (2004): 73–85. doi:10.1300/J122v25n01\_06.
- Schneider, Sandra L., Kirsten K. Ness, Sara Rockwell, Kelly Shaver, and Randy Brutkiewicz. *2012 Faculty Workload Survey*. Research report. Federal Demonstration Partnership, April 2014. [http://web.archive.org/web/20151022202705/http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga\\_087667.pdf](http://web.archive.org/web/20151022202705/http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_087667.pdf).
- Schwarzman, Alexander. "Supplemental Information: Who's Doing What and Why." PowerPoint presented at the CSE 2012 Annual Meeting, Seattle, WA, May 20, 2012. [http://web.archive.org/web/20151013195224/http://www.niso.org/apps/group\\_public/document.php?document\\_id=8591&wg\\_abbrev=supptechnical](http://web.archive.org/web/20151013195224/http://www.niso.org/apps/group_public/document.php?document_id=8591&wg_abbrev=supptechnical).
- . "Supplemental Materials Survey." *Information Standards Quarterly* 22, no. 3 (Summer 2010): 23–26. [http://web.archive.org/web/20151013194617/http://www.niso.org/apps/group\\_public/download.php/4886/Schwarzman\\_SuppMatlsSurvey\\_ISQ\\_v22no3.pdf](http://web.archive.org/web/20151013194617/http://www.niso.org/apps/group_public/download.php/4886/Schwarzman_SuppMatlsSurvey_ISQ_v22no3.pdf).
- Siebert, Sabina, Laura M. Machesky, and Robert H. Insall. "Overflow in Science and Its Implications for Trust." *eLife* 4 (2015): e10825. doi:10.7554/eLife.10825.
- Singh, Ajai R., and Shakuntala A Singh. "Ethical Obligation towards Research Subjects." *Mens Sana Monographs* 5, no. 1 (2007): 107–12. doi:10.4103/0973-1229.32153.

- Stewardship Gap Project. "The Problem." 2015. [http://web.archive.org/web/20160214022939/http://www.colorado.edu/ibs/cupc/stewardship\\_gap/](http://web.archive.org/web/20160214022939/http://www.colorado.edu/ibs/cupc/stewardship_gap/).
- Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. "Data Sharing by Scientists: Practices and Perceptions." *PLoS ONE* 6, no. 6 (2011): e21101. doi:10.1371/journal.pone.0021101.
- Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, and Kristina Dorsett. "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide." *PLoS ONE* 10, no. 8 (2015): e0134826. doi:10.1371/journal.pone.0134826.
- US Geological Survey. "Scientific Data Management Foundation." U.S. Geological Survey Instructional Memorandum No. IM OSQI 2015-01. US Geological Survey Manual, February 19, 2015. <http://web.archive.org/web/20151031000659/http://www.usgs.gov/usgs-manual/im/IM-OSQI-2015-01.html>.
- Vale, Ronald D. "Accelerating Scientific Publication in Biology." Preprint. bioRxiv, September 12, 2015. doi:10.1101/022368.
- . "Accelerating Scientific Publication in Biology." *Proceedings of the National Academy of Sciences* 112 (2015): 13,439–46. doi:10.1073/pnas.1511912112.
- Vaux, David L. "Research Methods: Know When Your Numbers Are Significant." *Nature* 492 (2012): 180–81. doi:10.1038/492180a.
- Wald, Chelsea. "Redefining Tenure at Medical Schools." *Science Careers*, March 6, 2009. doi:10.1126/science.caredit.a0900032.
- Wicherts, Jelte M., Denny Borsboom, Judith Kats, and Dylan Molenaar. "The Poor Availability of Psychological Research Data for Reanalysis." *American Psychologist* 61, no. 7 (October 2006): 726–28. doi:10.1037/0003-066X.61.7.726.
- Williams, Sarah C. "Data Sharing Interviews with Crop Sciences Faculty: Why They Share Data and How the Library Can Help." *Issues in Science and Technology Librarianship*, Spring 2013. doi:10.5062/F4T151M8.



**PART II**  
**Data Curation Services  
in Action**





## CHAPTER 6\*

# Research Data Services Maturity in Academic Libraries

*Inna Kouper, Kathleen Fear, Mayu Ishida,  
Christine Kollen, and Sarah C. Williams*

## Introduction

In 2012 only a small number of academic libraries offered research data services (RDS), but many were planning to do so within the next two years.<sup>1</sup> By 2013, 74 percent of respondents to an Association of Research Libraries (ARL) survey offered RDS, and an additional 23 percent were planning to do so.<sup>2</sup> Stimulated by shifts toward computational paradigms and the issuance of federal mandates to increase access to products of federally funded research, academic libraries recognize that the landscape of services changes quickly and that they need to support the changing needs of research and instruction.

To provide effective support for their constituencies, libraries must be proactive and develop services that look forward and yet accommodate the existing human, technological, and intellectual resources accumulated over the decades. Setting the stage for data curation in libraries means creating visionary approaches that supersede institutional differences while still enabling diversity in implementation. How do academic libraries approach data curation? What constitutes

---

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

an established RDS suite in an academic library? What can help in RDS evaluation, comparison, and improvement?

This chapter sets data curation in academic libraries within the broader context of RDS development and combines a historical overview of RDS thinking and implementations with an empirical analysis of libraries' RDS goals and activities. Using historical and current empirical data, the chapter synthesizes the state of RDS across academic libraries and argues that curation needs to be seen as part of a larger suite of services offered by libraries in support of the research life cycle and that the services evolve over time. To better understand this evolution and compare RDS across institutions, the chapter offers an empirically based framework of RDS maturity. A set of recommendations that libraries might consider to advance their RDS to the next maturity level is provided at the end.

## Research Data and Libraries

Since the 1950s, if not earlier, much of the work around data has been done by research communities as they grappled with global, inter-institutional data management and archiving.<sup>3</sup> North American academic libraries have also worked toward establishing research data services, though their services have often been anchored within their institutions. These early library data services were prominent in the areas of social science and GIS data reference and acquisition, but also in stewardship and sharing of data.<sup>4</sup> Conversations about data stewardship and the library's role in it tended to focus on needs within the university community. Thus, in 1965 I. de Sola Pool argued that

The storing of basic data in retrievable and manipulable form is, indeed, a library function. The library is an archive of that type of information that is of interest to many members of the university community and that is too bulky or expensive for each to retain or own. Each member of the faculty owns some books, but no member of the faculty can afford all the books he needs. The library provides the economy of shared-book usage. If this is a function of the library in the university, then clearly data archives also belong in the library....

Obviously, many data collections are so bulky or so expensive or so private that not even a university library can hope to own them. That, however, only suggests that specialization, division of labor, and linkage among libraries in a total library system are necessary in this field, as in other fields.<sup>5</sup>

The discussions of the 1970s and 1980s focused on staffing, institutional support, and computerized services to digitize and assist with machine-readable data.<sup>6</sup> The services of early data facilities already included acquisition, preservation, data cleaning, metadata, access and retrieval, reference, and data citation.<sup>7</sup> At the same time, libraries played a smaller role; among the forty-eight data-sharing facilities in the North America listed by Clubb et al. in 1985, thirty-one were associated with universities, with most of those facilities operating as collaborations between research and computing centers and sometimes libraries.<sup>8</sup> The Social Science Data and Program Library Service (DPLS) at the University of Wisconsin-Madison, for example, was established primarily by the faculty and could not be absorbed by the library because library staff at the time were not skilled in computers and data.<sup>9</sup>

In the late 1990s–2000s, with digital data and new forms of research on the rise, discussions shifted towards e-science, cyberinfrastructure, and digital curation, stimulated particularly by several seminal reports from the United States and the United Kingdom.<sup>10</sup> ARL recognized the importance of building members' awareness of the changes coming with the emergence of e-science and identified policies, skillful workforce, and research infrastructure as the primary areas of library engagement.<sup>11</sup> Data services have also been organized into tiers or areas that libraries could use to determine their current state, identify service gaps, and set goals and priorities.<sup>12</sup> Guidance on the development of data curation services “downstream” and “upstream” in the research life cycle was another way to define libraries' roles with RDS.<sup>13</sup>

A number of studies that examined the state and development of RDS in academic libraries show a clear trend of more academic libraries providing a broader range of e-science support and data-related services. In 2010, among 57 ARL libraries surveyed, 21 (37%) reported providing infrastructure or support services for e-science, with the rest being in the planning or no support stages.<sup>14</sup> Many libraries offered such services as information dissemination, consultations, and reference, as well as technology support (e.g., storage or software). A few libraries mentioned providing curation, stewardship, and preservation services. The common pressure points among the libraries included staffing and lack of infrastructure to handle, preserve, and provide access to data.

In 2012, about 44 percent of academic libraries surveyed provided reference support for finding data, and 20 percent or less provided other types of data-related services.<sup>15</sup> The services offered were predominantly in the informational or consultative category, such services as outreach and collaboration, training, and consultations. Creating web guides to help users find data and relevant information was one of the most common types of RDS among academic libraries. A rather rare category of technical or hands-on RDS included creating metadata and preparing, identifying, and deaccessioning data. The report also found that institutions with external funding were more likely to be involved in RDS de-



velopment, suggesting that funding agency requirements were driving the need for RDS.

By 2013, 74 percent or 54 of the ARL respondents offered RDS,<sup>16</sup> with many of them providing guidance and assistance with data management plans (DMPs). Three challenges identified in the ARL survey were (1) hiring and retraining staff, (2) building technical infrastructure, and (3) reaching out and collaborating with other stakeholders on campus. Research data management has been argued to be a major change in most librarians' responsibilities, as "data require different structural metadata, schemas, and vocabularies. Librarians who have adapted their skills are difficult to find."<sup>17</sup> ARL institutions approached RDS issues in diverse ways, and it was predicted that RDS would evolve over the next several years,<sup>18</sup> depending on institutional and funder policies as well as on financial and human resources available.

## The Current Landscape

To map the current landscape of RDS in academic libraries, we conducted a study of the 124 ARL libraries (as of September 2015) as those most likely to have started providing or planning for RDS. The study included content analysis of library webpages and a series of interviews with library administrators and program leads that examined their views of RDS goals, activities, and evolution. For content analysis we identified data-related webpages on library websites and coded their content for (1) the presence or absence of references to local repositories and to librarians dedicated to RDS, and (2) the presence or absence of references to particular types of services. The interviews were recorded and examined for common themes and specifics of RDS implementations. The results from both content analysis and interviews were used in a synthesizing depiction of the current landscape.

About half of the libraries (52%) indicated that they have a dedicated RDS position or librarian role on staff. The nature of dedicated positions varied from single librarians leading data services, to liaison librarians taking on research data management consultations, to full units or departments with multiple data consultants or specialists. This variety is consistent with earlier studies that found a range of staffing models and diverse position titles.<sup>19</sup>

The typology of services was developed using categories from the literature as well as from our own study.<sup>20</sup> The typology distills the surveyed libraries' service offerings into their core functional areas, such as "consultation and instruction," "collaboration and engagement," or "archiving and preservation" (see also appendix 6A for details on typology). Identifying core functional areas among varying implementations enabled us to consistently compare services across institutions and count their frequencies (see table 6.1).

**TABLE 6.1**  
**Research Data Services in the ARL Libraries**

<b>Group<sup>a</sup></b>	<b>Type of Service</b>	<b>% Libraries Mentioning Service on Website (N = 124)</b>
<i>Basic</i>	DMP assistance and mandate support	74%
	Consultations and instruction	73%
	Best practices and information dissemination	72%
<i>Intermediate</i>	Data deposit and repositories	49%
	Archiving and preservation	42%
	Collaboration and engagement	31%
	Metadata	30%
	Storage	27%
	Sharing and reuse	27%
<i>Advanced</i>	Data and researcher IDs	14%
	Data processing and analysis	13%
	Data curation	12%
	Acquisition	11%
	Copyright and ethics	10%
	Software and hardware	10%
	Data citation	10%
	Policies	7%
	Data reference	6%

a. Grouping is based on the frequency of service occurrence in the libraries, see more at the end of this section.

According to the webpages, most libraries (74%) provide DMP assistance and mandate support, including links to the DMPTool, an online service that contains DMP templates and allows researchers to create DMPs according to the funding agency requirements. Consultation and instruction as well as best practices and information dissemination are two other most frequent types of services (73% and 72%). Such capacity and partnership building is often mediated by subject librarians who are learning data management issues relevant to their disciplines and are ready to offer guidance on data management requirements for particular funding agencies.

The services of data deposit, archiving and preservation, collaboration and engagement, metadata, storage, and sharing and reuse were mentioned on fewer

webpages, ranging from 49 percent to 27 percent. These services require a higher level of institutional engagement and more financial, technological, and human resources. At the same time, developing a repository for data, or, more frequently, adapting an existing institutional repository to accept data, is a common second step for libraries offering data services. Thus, several of our respondents noted that they plan to pilot repository software and explore consortial options for data archiving. Despite only 49 percent of the libraries referring to data deposit as a service, many more (70%) had a repository that enabled data deposits. As data deposit requires efforts that are related to archiving and preservation, data and researcher IDs, and data curation, the beginnings of such services could have been considered part of many RDS efforts. Nevertheless, oftentimes such services were not specified as areas of concerted effort, and activities of deposit and preservation were used interchangeably.

A number of services were offered in less than 15 percent of the libraries, including permanent IDs for data and researchers, data curation, data processing and analysis, software and hardware support, data reference, and data citation. These kinds of services often depend on the specific user needs; additionally, they require a higher level of skill and expertise on the part of the library staff who offer them. A data reference librarian, for example, can be expected to be familiar with statistical software such as SPSS and understand how to manipulate numerical data in such software.\*

A striking difference in preservation efforts (42%) and curation efforts (12%) can probably be attributed to the differences in terminologies that various libraries employ to describe their efforts as well as to the awareness of the fuller spectrum of data services. At earlier stages of RDS, the terms “*preservation*” and “*curation*”, for example, can be used interchangeably. At more advanced stages of RDS, terminology becomes more specific because it refers to specific goals, tasks and responsibilities within a library. While the services of storage, archiving and preservation, and curation are connected and dependent on each other,<sup>20</sup> they become differentiated and sometimes specialized due to unique partnerships with IT units and commercial services.

Services that were the least common across libraries included support for copyright and ethics, software and hardware, data citation, and policy development. These areas are among the most challenging to implement in the libraries, as stakeholders in data exchanges—including producers, providers, publishers, and consumers—are trying to understand the best ways to ensure open sharing while protecting ownership and to create tools for storing, analyzing, and sharing data at scale. Many respondents in our study confirmed that some work on devel-

---

\* See, for example, a data reference librarian job description: “Data Reference Librarian,” job opening at Harvard College Library, posted to IASSIST August 20, 2008, <http://www.iassistdata.org/resources/jobs/1612>.

oping data policies was being done, but it involved university-wide consultations and collaborations with institutional review boards, research administration, and information technology units. Some libraries, while acknowledging the need for data policies to guide their service provisioning and to enable data sharing, postpone such work as it needs to be consistent with the funding mandates, publishing policies, and other areas that involve data. The early work on data policies includes efforts to incorporate data management into institutional research policies and to increase awareness of the existing policies with regard to sensitive data and data ownership within universities.

To provide an additional way of comparing RDS across academic libraries and to build the foundation for the discussion about RDS maturity below, the typology of services is further grouped into three categories based on the *frequency of service occurrence* in the libraries: the **basic** services group includes services that exist in more than 50 percent of the libraries, the **intermediate** services group includes services that exist in less than 50 percent but more than 15 percent of the libraries, and the **advanced** services group includes services that exist in less than 15 percent of the libraries. While frequency alone cannot be an indicator of RDS maturity, such an approach has found support in our interviews and in the literature. Respondents in our interviews reflected that DMP services were typically the first type of services they offered when starting RDS at their institutions, while also noting that they needed to move beyond that and basic policy compliance and informational services. Similarly, Fearon noted that many libraries started their RDS with support for DMPs, with almost 90 percent of the libraries providing DMP support and consultation services.<sup>22</sup> The basic group of services naturally lends itself to the beginning stages of RDS development as it is a straightforward outgrowth of the work librarians do in advisory and reference services and is relatively easy to implement; the intermediate and advanced groups require more skills, better stakeholder engagement and institutional support, and more resources.

## RDS Maturity

In the previous section, we introduced a typology of data services and, based on our content analysis and interviews, posited that the most frequently offered services are those that represent a straightforward entry point into RDS, while those that are more challenging—more resource-intensive, more specialized, and more reliant on institutional support—are both rarer and more advanced. In the following section, we develop this initial statement into a maturity model for RDS.

Maturity evaluation is a common approach to determining the level of sophistication of services or products. One of the earlier, better known examples of such models, the Capability Maturity Model for Software (CMM-SW), was

developed in the 1990s to aid the US Department of Defense in software acquisition.<sup>23</sup> The model's goals were to appraise software processes and help organizations to move from chaotic ad hoc processes of development to disciplined and optimal ones.<sup>24</sup> The model developers distinguished between immature and mature software organizations and argued that the former are primarily reactionary and focus on solving immediate problems, while the latter are based on solid management techniques, such as consistent planning, communication, pilot testing, cost-benefit analysis, and defined roles and responsibilities.

Recently, Qin, Crowston, Flynn, and Kirkland proposed using maturity levels similar to the CMM-SW to assess and improve research data management (RDM) practices in research projects.<sup>25</sup> They described the five levels in application to RDM as follows. The first, *initial* level of RDM relies on competent individuals and heroic efforts, making the data management efforts unreliable. The second, *managed* level of RDM is based on the procedures and policies established in advance for each project, which makes it difficult to apply RDM across projects. The third, *defined* level is characterized by established and repeatable procedures that can be used across projects. The fourth, *quantitatively managed* level adds metrics that help to evaluate processes and progress. The final, *optimizing* level focuses on improvement and identification of weaknesses and inefficiencies that can be addressed proactively. The maturity levels are suggested to be applied to the following key process and practice areas: (1) data management in general; (2) data acquisition, processing, and quality assurance; (3) data description and representation; (4) data dissemination; and (5) repository services and preservation.

The capability maturity framework guide for data management proposed by the Australian National Data Service (ANDS) uses the same maturity levels as CMM-SW and CMM RDM, but it identifies different process areas: (1) institutional policies and procedures; (2) IT infrastructure; (3) support services; and (4) managing metadata.<sup>26</sup> For each of the areas, the processes move from being ad hoc and disorganized to being defined, standardized, managed, and optimized. Yet, there is one major difference. The CMM-RDM framework fits with the research life cycle approach and, with data management, can be applied to the stages of data collection, processing, dissemination, and preservation and, therefore, can be applied at the project level. On the other hand, the process areas of the ANDS model identify larger areas within the institutional context (e.g., policies, infrastructure, education, and metadata) that need to be in place before data management within the life cycle can take place.

These models, and many other capability models that have been developed over the last few decades,<sup>27</sup> provide guidance in terms of the trajectory that a team, a project, a service, or an organization can go through to become a well-managed unit with clear goals and path toward deliverable results. At the same time, the models offer rather loose definitions of each level and leave it up to the user of the model to determine whether the processes within an organization are sufficient-

ly organized, documented, managed, or optimized. CMM-RDM provides more guidance, but it is an outward looking model; that is, it guides the development of data management for data management “consumers,” such as researchers or data managers, rather than librarians. It is also not clear how much empirical ground-work went into the process areas development and maturity levels. An “inward” approach to maturity modeling that looks at data management “providers,” or organizations supporting research in academic institutions, will better suit the needs of research data services being developed and evaluated in academic libraries.

Similar to the maturity of software development or data management, RDS maturity can be defined as the extent to which specific services are defined, managed, and evaluated in their impact and effectiveness. Each service and the system of services as a whole can be evaluated in its richness and consistency with the overall organizational goals. To be well-developed and well-understood throughout an organization, RDS need to rely on dissemination and training, and constant user feedback. Maturity also implies consistent growth and improvement via a disciplined and optimized approach.

The difference between software development and RDS is in how growth over time and improvements are conceptualized. In the context of software development, the goal is to improve processes in order to more quickly, reliably, and effectively turn out new products, often in a competitive market environment. For academic libraries, however, there is a complex interaction between the goals of RDS and the bigger strategic goals of the library and the institution; further, individual institutions’ RDS efforts are just one part of a complex and largely cooperative network of data support, which includes external entities such as disciplinary and other repositories, funders and their initiatives, commercial services, and so on. As a result, the highest, optimized level of maturity may have a different meaning for various institutions depending on their missions and goals. Knowing where the “finish line” is in terms of the nature of services provided in a particular institutional context is as important as knowing what services to implement.

The key areas and levels proposed in the maturity model in table 6.2 are based on our empirical analysis of the ARL libraries, particularly on the analysis of interviews with library administrators and program leads regarding their views on immediate RDS implementation directions, short-term goals, and future plans. While analyzing the interviews and extracting common themes and approaches, we found that many interviewees agreed that in order to develop strong and mature RDS, a library needs to have the following: a mission that is consistent with the institutional mission, services that match user needs, qualified and dedicated staff, strong relationships with other units on campus and with other institutions, and established policies that guide data collection, sharing, and use. The synthesis of these themes along with many other discussions mentioned above formed the basis of eight key areas of maturity: leadership, services, users and stakeholders, research life cycle support, governance, cost and budgeting, cross-unit collaboration, and human capital.

**TABLE 6.2**  
**Research Data Services Maturity Model**

<b>Maturity Levels</b> <b>Key Areas</b>	<b>Basic :: Foundation Building</b>	<b>Intermediate :: Organization and Standardization</b>	<b>Advanced :: Monitoring and Optimization</b>
<i>Leadership (vision, strategy, culture)</i>	Response to mandates and external activities	Data strategies are coordinated with institutional strategic documents.	Data strategies guide service development and assessment.
<i>Services</i>	DMP assistance, consultations and instruction, best practices and information dissemination	Data deposit and repositories, archiving and preservation, collaboration and engagement, metadata, storage, data sharing and reuse	Permanent IDs for data and researchers, data curation, data processing and analysis, software and hardware, data citation
<i>Users and stakeholders</i>	Addressing individual requests	User strategy is based on needs assessment.	User needs are regularly evaluated, and services and needs shape each other.
<i>Research life cycle support</i>	Support on one end (upstream with DMP or downstream with data deposit)	Support broadens and formalizes for both upstream and downstream.	Support is embedded in the life cycle.
<i>Governance</i>	No policies, or reliance on institutional policies	Data mentioned in other policies or one general data policy	Set of policies from acquisition to storage to curation and dissemination
<i>Cost and budgeting</i>	Spending is a burden; each data-related expense needs to be requested and justified.	Spending brings benefits and creates opportunities.	Budgeting for growth and sustainability
<i>Cross-unit collaboration</i>	None, or ad hoc meetings and committees within institution	Joint initiatives with other units	Formal partnerships within and outside, support from university administration
<i>Human capital</i>	Other staff, such as subject librarians, assume data responsibilities, ad hoc training	Solo librarian or a working group, consistent professional training	Dedicated team with shared or specialized responsibilities, strong skills, continuous learning



The RDS maturity levels are simplified from five to three as compared to other CMMs to aid in clearer definitions and subsequent validation effort. The three levels also effectively represent the diversity of RDS approaches among the academic libraries in our study, which corresponded to the basic, intermediate, and advanced categorization and converged on the following three stages: (1) foundation building, (2) organization and standardization, and (3) monitoring and optimization.

During foundation building, the library focuses on implementing services that do not require significant resources and expertise, and it is done with limited staff support. The implementation efforts are mostly driven by mandates and individual user requests, and no significant cross-unit collaboration and user assessment exists. Each data-related expense needs to be justified because it potentially takes away from other library activities.

At the level of organization and standardization, the library gets involved in strategic efforts to coordinate its activities with the institutional goals and mission. The leadership becomes less reactive and more focused on a stronger view of the future and the role the services will play in shaping it. The services are customized to meet institution-specific requirements; they are based on user needs assessment and cross-unit collaboration. Professional development becomes part of the library activities, and spending becomes more organized to spur further development.

At the monitoring and optimization level, services become more diverse and become embedded in the research life cycle. The library not only engages users and stakeholders and understands their needs, but also develops an effective feedback system. The library also develops a comprehensive set of policies and strategic documents and builds formal external and internal cross-unit partnerships. The data services team structure and organization moves from solo librarians to dedicated, multifunctional, or specialized teams.<sup>28</sup>

## Looking into the Future

As academic libraries continue to grow their RDS programs, there are two areas of strategic activities that are of primary importance in developing appropriately targeted, effective services. First, libraries need to continue to assess what their peer institutions currently offer and ask: How similar and different they are? What they are trying to achieve? What they have learned and would do differently? and, more importantly, Why they are offering those particular services? Second, libraries should also aim for service development that is not simply reactive; developing a vision for RDS is a critical precursor to selecting impactful services to implement. This study provides a baseline that can be used to trace RDS development and improvements across institutions as well as a model for evaluating and building RDS programs.

A key take-away from this study is that more advanced services are probably those that are most closely targeted to the needs of individual institutions' communities but are also cognizant of the broader research communities to which individuals belong. The institutional approach is one way to address RDS needs, and academic libraries are playing an important role in the national and global data ecosystem.<sup>29</sup> More mature RDS programs are not necessarily those that offer the longest menu of services or employ the largest number of staff, but rather those whose activities are more deeply embedded in the mission and activities of the library and the broader institution. Mature RDS services have strong connections within and outside the library, a plan for sustainability in place, well-developed policies, and so on. In other words, a mature RDS program is one where services are chosen carefully, and then carefully organized, monitored, and optimized.

To some extent, high levels of maturity reflect a high level of organizational buy-in: a sustainable budget for RDS, for example, is not something that can be accomplished in isolation. Our maturity model for RDS serves a dual purpose; it is a useful tool for identifying gaps and setting priorities, but it can also be a valuable tool for communicating with library administration. Part of developing RDS is asking for resources and support from the library, which means it is important not only to express the goals and vision for RDS specifically, but also to align them with the broader strategic goals and vision of the library and the institution. Many respondents in our interviews acknowledged resource limitations and recognized the importance of such an alignment.

Opportunities abound in building RDS. For libraries looking to take the next step with their services, it is critical to determine which opportunities are aligned with their priorities, whether it is developing a new service, building partnerships, or planning for assessment of existing services. Looking at what services peers offer as well as self-assessing a library's current RDS maturity level helps to sort out which opportunities will provide the most value in the long run.

## Appendix 6A: Typology of Services and Their Descriptions on Websites

<b>Type of Service</b>	<b>Explanation</b>
<i>Acquisition</i>	Statements that describe acquisition and collection management with regard to data
<i>Archiving and preservation</i>	Statements that describe long-term archiving and preservation of data
<i>Best practices and information dissemination</i>	Statements that describe efforts to collect and disseminate information about (best) practices in data management and sharing, mostly via websites and other similar types of materials
<i>Collaboration and engagement</i>	Statements that describe efforts to engage with faculty, other units on campus, or other organizations
<i>Consultations and instruction</i>	Statements that describe consultation and instruction initiatives, including workshops, seminars, and so on (more active orientation than dissemination)
<i>Copyright and ethics</i>	Statements that describe efforts to providing information about intellectual property and ethical uses of data
<i>Data processing and analysis</i>	Statements that describe assistance and guidance on data processing and analysis resources and issues
<i>Data and researcher IDs</i>	Statements about services that help to create and maintain permanent identification for people and documents
<i>Data citation</i>	Statements about guidance on how and why to cite data
<i>Data curation</i>	Statements that describe activities of curation with regard to data
<i>Data deposit and repositories</i>	Statements that describe assistance in finding and using appropriate repositories (disciplinary or institutional)
<i>Data reference</i>	Statements about reference-type services, including search, sources, and use of tools
<i>DMP assistance and mandate support</i>	Statements about assistance with DMPs and compliance with funding agencies mandates
<i>Metadata</i>	Statements about assistance with generating or structuring metadata
<i>Policies</i>	Statements about creating, developing, or providing policies with regard to data
<i>Sharing and reuse</i>	Statements that describe support of sharing and reuse
<i>Software and hardware</i>	Statements that describe efforts to provide or inform about hardware and software resources to process and analyze data
<i>Storage</i>	Statements that describe efforts to provide short-term and long-term storage for data

## Notes

1. Carol Tenopir, Ben Birch, and Suzie Allard, *Academic Libraries and Research Data Services* (Chicago: Association of College and Research Libraries, 2012), [http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir\\_Birch\\_Allard.pdf](http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf).
2. David Fearon Jr., Betsy Gunia, Sherry Lake, Barbara E. Pralle, and Andrew L. Sallans, *Research Data Management Services: SPEC Kit 334*: (Washington, DC: Association of Research Libraries, July 2013), <http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/>.
3. Mustapha Mokrane, "Global Data for Global Science: The New ICSU World Data System," *IAHS Newsletter*, April 2013, 4, <http://iahs.info/uploads/dms/16101.IAHS%20Newsletter%20105%20-%20Final2.pdf>; Inter-university Consortium for Political and Social Research, "About ICPSR," accessed February 3, 2016, <https://www.icpsr.umich.edu/icpsrweb/content/membership/about.html>.
4. Anna Gold, "Libraries and the Data Challenge: Roles and Actions for Libraries," *D-Lib Magazine* 13, no. 9/10 (2007), <http://www.dlib.org/dlib/september07/gold/09gold-pt2.html>.
5. Ithiel de Sola Pool, "Data Archives and Libraries," In *INTREX: Report of a Planning Conference on Information Transfer Experiments*, ed. Carl F. J. Overhage and R. Joyce Harman (Woods Hole, MA: MIT Press, 1965), 179–80.
6. Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, eds. *Sharing Research Data* (Washington, DC: National Academies Press, 1985).
7. Laine Ruus, "The University of British Columbia Data Library: An Overview," *Library Trends* 30, no. 3 (1982): 397–407.
8. Jerome M. Clubb, Erik W. Austin, Carolyn L. Geda, and Michael W. Traugott, "Sharing Research Data in the Social Sciences," In *Sharing Research Data*, eds. Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf (Washington, DC: National Academies Press, 1985), 77–79.
9. Alice Robbin, "The Data and Program Library Service: A Case Study in Organizing Special Libraries for Computer-Readable Statistical Data," *Library Trends* 30, no. 3 (1982): 407–32.
10. Interagency Working Group on Digital Data, *Harnessing the Power of Digital Data for Science and Society* (Arlington, VA: The Networking and Information Technology Research and Development Program, 2009), [https://www.nitrd.gov/About/Harnessing\\_Power\\_Web.pdf](https://www.nitrd.gov/About/Harnessing_Power_Web.pdf); Elizabeth Yakel, "Digital Curation," *OCLC Systems and Services* 23, no. 4 (2007): 335–40, doi:10.1108/10650750710831466; Cyberinfrastructure Council, *Cyberinfrastructure Vision for 21st Century Discovery*, NSF 07-28 (Arlington, VA: National Science Foundation, 2007), <http://www.nsf.gov/pubs/2007/nsf0728/>; Office of Science and Innovation e-Infrastructure Working Group, *Developing the UK's e-Infrastructure for Science and Innovation* (National e-Science Centre, 2007), <http://www.nesc.ac.uk/documents/OSI/report.pdf>.
11. Joint Task Force on Library Support for E-Science, *Agenda for Developing e-Science in Research Libraries* (Washington, DC: Association of Research Libraries, 2007), <http://www.arl.org/storage/documents/publications/escience-report-final-2007.pdf>.
12. Rebecca Reznik-Zellen, Jessica Adamick, and Stephen McGinty, "Tiers of Research Data Support Services," *Journal of eScience Librarianship* 1, no. 1 (2012): 27–35,

- doi:10.7191/jeslib.2012.1002; Kathleen Shearer, and Diego Argaez, *Addressing the Research Data Gap* (Ottawa: Canadian Association of Research Libraries, 2010), [http://www.carl-abrc.ca/uploads/pdfs/library\\_roles-final.pdf](http://www.carl-abrc.ca/uploads/pdfs/library_roles-final.pdf).
13. Gold, "Libraries and the Data Challenge"; Anna Gold, *Data Curation and Libraries* (San Luis Obispo, CA: California Polytechnic State University Office of the Dean [Library], 2010), <http://works.bepress.com/agold019/>.
  14. Catherine Soehner, Catherine Steeves, and Jennifer Ward, *E-Science and Data Support Services* (Washington, DC: Association of Research Libraries, 2010), <http://www.arl.org/storage/documents/publications/escience-report-2010.pdf>.
  15. Tenopir, Birch, and Allard, *Academic Libraries and Research Data Services*.
  16. Fearon et al., *Research Data Management Services*.
  17. P. Bryan Heidorn, "The Emerging Role of Libraries in Data Curation and E-sciences," *Journal of Library Administration* 51, no. 7–8 (2011): 670, doi:10.1080/01930826.2011.601269.
  18. Fearon et al., *Research Data Management Services*, 20.
  19. *Ibid.*, 17–18; Katherine G. Akers, Fe C. Sferdean, Natsuko H. Nicholls, and Jennifer A. Green, "Building Support for Research Data Management: Biographies of Eight Research Universities," *International Journal of Digital Curation* 9, no. 4 (2014): 171–91, doi:10.2218/ijdc.v9i2.327.
  20. Inna Kouper, Kathleen Fear, Mayu Ishida, Christine Kollen, and Sarah Williams, "Research Data Services Vision(s): An Analysis of North American Research Libraries" (presentation at the 41st IASSIST Annual Conference, Minneapolis, MN, June 2–5, 2015). <http://www.slideshare.net/InnaKouper/research-data-services-visions-an-analysis-of-north-american-research-libraries>.
  21. G. Sayeed Choudhury, Carole L. Palmer, Karen S. Baker, and Timothy DiLauro, "Levels of Services and Curation for High Functioning Data" (presentation at the 8th International Digital Curation Conference, Amsterdam, the Netherlands, January 14–17, 2013), <http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster192.pdf>.
  22. Fearon et al., *Research Data Management Services*, 13.
  23. Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber, *Capability Maturity Model for Software, Version 1.1* (Pittsburgh, PA: Carnegie Mellon University Software Engineering Institute, 1993), [https://resources.sei.cmu.edu/asset\\_files/TechnicalReport/1993\\_005\\_001\\_16211.pdf](https://resources.sei.cmu.edu/asset_files/TechnicalReport/1993_005_001_16211.pdf).
  24. James Herbsleb, David Zubrow, Dennis Goldenson, Will Hayes, and Mark Paulk, "Software Quality and the Capability Maturity Model," *Communications of the ACM* 40, no. 6 (1997): 30–40, doi:10.1145/255656.255692.
  25. Jian Qin, Kevin Crowston, Charlotte Flynn, and Arden Kirkland, *Development and Dissemination of a Capability Maturity Model for Research Data Management Training and Performance Assessment: A Final Report to the Interuniversity Consortium for Political and Social Research*, School of Information Studies, Syracuse University, accessed November 15, 2015, <http://datacommunity.icpsr.umich.edu/sites/default/files/CMM%20for%20RDM%20project%20report%20package.pdf>.
  26. Australian National Data Services, "Research Data Management Framework: Capability Maturity Guide," accessed November 13, 2015, [http://www.ands.org.au/\\_data/assets/image/0004/384754/capability-maturity-table.JPG](http://www.ands.org.au/_data/assets/image/0004/384754/capability-maturity-table.JPG).
  27. Stella Fowler, "JISC Managing Research Data Project Maturity Model," University of the West of England, February 23, 2012, accessed February 2, 2016, <http://www2.uwe>.

- ac.uk/services/library/using\_the\_library/Services%20for%20researchers/maturity-model-v.1.1.pdf; Wim Hugo, "A Maturity Model for Digital Data Centers," *Data Science Journal*, no. 12 (2013): WDS189–WDS192, doi:10.2481/dsj.WDS-032.
28. Cheryl Thompson, Charles Humphrey, and Michael Witt, "Exploring Organizational Approaches to Research Data in Academic Libraries: Which Archetype Fits Your Library?" (presentation at the 6th Research Data Alliance Plenary Meeting, Paris, France, September 23–25, 2015). [http://www.slideshare.net/RDA\\_Data\\_Share/exploring-organizational-approaches-to-research-data-in-academic-libraries-by-cheryl-thompson](http://www.slideshare.net/RDA_Data_Share/exploring-organizational-approaches-to-research-data-in-academic-libraries-by-cheryl-thompson).
  29. Compute Canada, "Compute Canada and the Canadian Association of Research Libraries Join Forces to Build a National Research Data Platform," accessed February 3, 2016. <https://www.computeCanada.ca/research/compute-canada-and-the-canadian-association-of-research-libraries-join-forces-to-build-a-national-research-data-platform/>.

## Bibliography

- Akers, Katherine G., Fe C. Sferdean, Natsuko H. Nicholls, and Jennifer A. Green. "Building Support for Research Data Management: Biographies of Eight Research Universities." *International Journal of Digital Curation* 9, no. 4 (2014): 171–91. doi:10.2218/ijdc.v9i2.327.
- Australian National Data Services. "Research Data Management Framework: Capability Maturity Guide." Accessed November 13, 2015, [http://www.ands.org.au/\\_\\_data/assets/image/0004/384754/capability-maturity-table.JPG](http://www.ands.org.au/__data/assets/image/0004/384754/capability-maturity-table.JPG).
- Choudhury, G. Sayeed, Carole L. Palmer, Karen S. Baker, and Timothy DiLauro, "Levels of Services and Curation for High Functioning Data" (presentation at the 8th International Digital Curation Conference, Amsterdam, the Netherlands, January 14–17, 2013), <http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster192.pdf>.
- Clubb, Jerome M., Erik W. Austin, Carolyn L. Geda, and Michael W. Traugott, "Sharing Research Data in the Social Sciences," In *Sharing Research Data*, edited by Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, 77–79. Washington, DC: National Academies Press, 1985.
- Compute Canada and the Canadian Association of Research Libraries Join Forces to Build a National Research Data Platform," accessed February 3, 2016. <https://www.computeCanada.ca/research/compute-canada-and-the-canadian-association-of-research-libraries-join-forces-to-build-a-national-research-data-platform/>
- Cyberinfrastructure Council. *Cyberinfrastructure Vision for 21st Century Discovery*. NSF 07-28. Arlington, VA: National Science Foundation, 2007. <http://www.nsf.gov/pubs/2007/nsf0728/>.
- de Sola Pool, Ithiel. "Data Archives and Libraries." In *INTREX: Report of a Planning Conference on Information Transfer Experiments*, edited by Carl F. J. Overhage and R. Joyce Harman, 175–81. Woods Hole, MA: MIT Press, 1965.
- Fearon, David Jr., Betsy Gunia, Sherry Lake, Barbara E. Pralle, and Andrew L. Sallans. *Research Data Management Services: SPEC Kit 334*. Washington, DC: Association of Research Libraries, July 2013. <http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/>.
- Fienberg, Stephen E., Margaret E. Martin, and Miron L. Straf, eds. *Sharing Research Data*. Washington, DC: National Academies Press, 1985.

- Fowler, Stella. "JISC Managing Research Data Project Maturity Model." University of the West of England, February 23, 2012, accessed February 2, 2016, [http://www2.uwe.ac.uk/services/library/using\\_the\\_library/Services%20for%20researchers/maturity-model-v.1.1.pdf](http://www2.uwe.ac.uk/services/library/using_the_library/Services%20for%20researchers/maturity-model-v.1.1.pdf).
- Gold, Anna. *Data Curation and Libraries: Short-Term Developments, Long-Term Prospects*. San Luis Obispo, CA: California Polytechnic State University Office of the Dean [Library], 2010. <http://works.bepress.com/agold01/9/>.
- . "Libraries and the Data Challenge: Roles and Actions for Libraries." *D-Lib Magazine* 13, no. 9/10 (2007). <http://www.dlib.org/dlib/september07/gold/09gold-pt2.html>.
- Heidorn, P. Bryan. "The Emerging Role of Libraries in Data Curation and E-science." *Journal of Library Administration* 51, no. 7–8 (2011): 662–72. doi:10.1080/01930826.2011.601269.
- Herbsleb, James, David Zubrow, Dennis Goldenson, Will Hayes, and Mark Paulk. "Software Quality and the Capability Maturity Model." *Communications of the ACM* 40, no. 6 (1997): 30–40. doi:10.1145/255656.255692.
- Hugo, Wim, "A Maturity Model for Digital Data Centers," *Data Science Journal*, no. 12 (2013): WDS189–WDS192, doi:10.2481/dsj.WDS-032.
- Interagency Working Group on Digital Data. *Harnessing the Power of Digital Data for Science and Society*. Arlington, VA: The Networking and Information Technology Research and Development Program, 2009. [https://www.nitrd.gov/About/Harnessing\\_Power\\_Web.pdf](https://www.nitrd.gov/About/Harnessing_Power_Web.pdf).
- Inter-university Consortium for Political and Social Research, "About ICPSR," accessed February 3, 2016, <https://www.icpsr.umich.edu/icpsrweb/content/membership/about.html>.
- Joint Task Force on Library Support for E-Science. *Agenda for Developing E-Science in Research Libraries*. Washington, DC: Association of Research Libraries, 2007. <http://www.arl.org/storage/documents/publications/escience-report-final-2007.pdf>.
- Kouper, Inna, Kathleen Fear, Mayu Ishida, Christine Kollen, and Sarah Williams. "Research Data Services Vision(s): An Analysis of North American Research Libraries." Presentation at the 41st IASSIST Annual Conference, Minneapolis, MN, June 2–5, 2015. <http://www.slideshare.net/InnaKouper/research-data-services-visions-an-analysis-of-north-american-research-libraries>.
- Mustapha Mokrane, "Global Data for Global Science: The New ICSU World Data System," *LAHS Newsletter*, April 2013, 4, <http://iahs.info/uploads/dms/16101.LAHS%20Newsletter%20105%20-%20Final2.pdf>.
- Office of Science and Innovation e-Infrastructure Working Group. *Developing the UK's e-Infrastructure for Science and Innovation*. National e-Science Centre, 2007. <http://www.nesc.ac.uk/documents/OSI/report.pdf>.
- Paulk, Mark C., Bill Curtis, Mary Beth Chrissis, and Charles V. Weber. *Capability Maturity Model for Software, Version 1.1*. Pittsburgh, PA: Carnegie Mellon University Software Engineering Institute, 1993. [https://resources.sei.cmu.edu/asset\\_files/TechnicalReport/1993\\_005\\_001\\_16211.pdf](https://resources.sei.cmu.edu/asset_files/TechnicalReport/1993_005_001_16211.pdf).
- Qin, Jian, Kevin Crowston, Charlotte Flynn, and Arden Kirkland. *Development and Dissemination of a Capability Maturity Model for Research Data Management Training and Performance Assessment: A Final Report to the Interuniversity Consortium for Political and Social Research*. School of Information Studies, Syracuse University. Accessed



- Nov. 15, 2015. <http://datacommunity.icpsr.umich.edu/sites/default/files/CMM%20for%20RDM%20project%20report%20package.pdf>.
- Reznik-Zellen, Rebecca, Jessica Adamick, and Stephen McGinty. "Tiers of Research Data Support Services." *Journal of eScience Librarianship* 1, no. 1 (2012): 27–35. doi:10.7191/jeslib.2012.1002.
- Robbin, Alice. "The Data and Program Library Service: A Case Study in Organizing Special Libraries for Computer-Readable Statistical Data." *Library Trends* 30, no. 3 (1982): 407–32.
- Ruus, Laine. "The University of British Columbia Data Library: An Overview." *Library Trends* 30, no. 3 (1982): 397–407.
- Shearer, Kathleen, and Diego Argaez. *Addressing the Research Data Gap: A Review of Novel Services for Libraries*. Ottawa, Ontario: Canadian Association of Research Libraries, 2010. <https://libshare.library.gatech.edu/docs/DOC-4280>.
- Soehner, Catherine, Catherine Steeves, and Jennifer Ward. *E-Science and Data Support Services: A Study of ARL Member Institutions*. Washington, DC: Association of Research Libraries, 2010. <http://www.arl.org/storage/documents/publications/escience-report-2010.pdf>.
- Tenopir, Carol, Ben Birch, and Suzie Allard. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*. Chicago: Association of College and Research Libraries, 2012. [http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir\\_Birch\\_Allard.pdf](http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf).
- Thompson, Cheryl, Charles Humphrey, and Michael Witt. "Exploring Organizational Approaches to Research Data in Academic Libraries: Which Archetype Fits Your Library?" Presentation at the 6th Research Data Alliance Plenary Meeting, Paris, France, September 23–25, 2015. [http://www.slideshare.net/RDA\\_Data\\_Share/exploring-organizational-approaches-to-research-data-in-academic-libraries-by-cher-yl-thompson](http://www.slideshare.net/RDA_Data_Share/exploring-organizational-approaches-to-research-data-in-academic-libraries-by-cher-yl-thompson).
- Yakel, Elizabeth. "Digital Curation." *OCLC Systems and Services* 23, no. 4 (2007): 335–40. doi:10.1108/10650750710831466.



CHAPTER 7\*

# Extending Data Curation Service Models for Academic Library and Institutional Repositories

*Jon Wheeler*

## Introduction

Development of research data management (RDM) and curation services remains both a priority and a challenge for many academic research libraries. Broadly speaking, while library service models continue to evolve to meet the data management needs of researchers accountable to emerging funder requirements, it remains true that many librarians seek clarification about their role in support of data curation.<sup>1</sup> Discussion by Antell and colleagues and Nielsen and Hjørland highlight in particular some of the contradictions librarians perceive between the drive to develop data management skills and the practicality of situating these skills within libraries generally.<sup>2</sup>

A similar contrast exists between the technical capabilities and expectations regarding the use of institutional repositories for data publication and preservation.

---

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

Because IRs may include content from multiple disciplines and a variety of types—for example electronic theses and dissertations (ETD) and research posters—their utility as data repositories may be legitimately called into question. As noted by Don MacMillan, IRs as data repositories may “further fragment the data landscape and may result in making data more difficult to find than it would be in larger subject-specific or interdisciplinary repositories.”<sup>3</sup> Even in cases where one may argue that an IR is a better-than-nothing option, issues described in McGovern and McKay and Jain illustrate how the diversity of IR service models can inhibit their utility when publication workflows are not based on best practices or otherwise modeled against disciplinary standards.<sup>4</sup> However, with such concerns in mind, a consideration of established library data management services, functions, and roles provides context for the description and development of an IR data service focused on archiving and mirroring collections previously published within domain repositories. Beginning with an overview of the suitability of IRs for this purpose, the chapter further addresses how such a service aligns with existing capabilities and provides illustrative scenarios and strategies for implementation.

## Conceptual Models and Rationale

Establishing IRs as mirrors of data collections held by domain repositories is a service capability described at least implicitly in the literature. In particular, the “web of repositories” model presented by Baker and Yarmey and elaborated upon by Baker and Millerand is relevant because of the model’s emphasis on situated, role-based services oriented toward data management within local and nonlocal contexts.<sup>5</sup> Locality and distance are here understood not as spatial or geographic constraints, but rather refer to a repository’s support for or contribution to data management at different stages in the research life cycle. Additionally, the model is understood to be nonlinear in the sense that data do not necessarily move in sequence from one repository setting to another, but may be hosted or mirrored across systems. As the context changes, so does the community served by the corresponding repository, which necessarily impacts the services provided to manage data in that context. A recent and innovative example of this model is provided by Walters, in which IRs are not preservation end points in themselves, but rather act as communication layers between production services and preservation architectures.<sup>6</sup>

The proposed mirroring service is further informed by the “preservation as relay” model described by Janée and colleagues.<sup>7</sup> Whereas the web of repositories model explicitly includes mirroring collections across multiple sites, the preservation as relay model more narrowly refers to a complete handoff or asset transfer wherein different types of repositories fulfill preservation requirements at different times. As an example, a short- or near-term repository may commit to providing services or taking necessary actions to transform, migrate, or curate data for five

years. In the event long-term support is required, the migration to another repository altogether, perhaps one with a five-to-ten-year remit, becomes a relay or a handoff. This model includes dark archiving, or periods during which no archive is able to provide public access to the data, with the expectation that appropriate preservation practices are in place to successfully recover the data when necessary.

While these two models provide a broad rationale for establishing IRs as complementary services to domain repositories, further justification for a mirroring service is provided by drivers including funder policies and the DR ecosystem. Policy-wise, the specification within funder data management plan (DMP) recommendations of “data archiving” or “data preservation” as distinct from “data sharing” strategies is relevant (see for example recommendations from the National Science Foundation and the Department of Energy<sup>8</sup>), as the practical difference between publishing and archiving—not to mention between archiving and backup—is not intuitive across disciplines. This is a significant issue, as a lack of distinction between these concepts can result in noncompliance and put data at risk. An illustrative example is provided by Choudhury, who relates how project team members from the Sloan Digital Sky Survey assumed that their data were sufficiently archived because they had been securely stored and backed up.<sup>9</sup> Even putting aside compliance concerns, the risk of data loss in such circumstances is further illustrated by Uhler’s concept of “information gulags” in which data are “preserved” within systems that are “highly distributed, silent, and invisible.”<sup>10</sup> Here the conflation of archival with sharing and backup processes contributes to a proliferation of these invisible data silos when systems and strategies are adopted that negatively impact the discoverability and usability of data. Establishing IRs as complementary archives of DR collections is one means of preventing information gulags by enlarging the context of discovery, accessibility, and exposure of data to users.

Finally, further practical justification is found among concerns about the sustainability and preservation-readiness of many DRs. As noted in the literature, coverage for data curation across the life cycle is well established within disciplines such as astronomy and certain subfields of biology.<sup>11</sup> However, the existence of established, trustworthy repositories across disciplines is the exception rather than the rule. This is a two-fold problem in that a given discipline may on the one hand lack established repositories, while on the other hand available repositories may not provide sufficient preservation support to satisfy funder expectations. For example, Castelli and colleagues enumerate multiple barriers that impact data discovery and preservation among data centers and research digital libraries.<sup>12</sup> In particular, that data may be documented only enough to support discovery or citation,<sup>13</sup> and the protocols in place for export or federation of resources may be limited or otherwise not based on best practices or standards such as OAI-PMH.<sup>14</sup> Sustainability concerns due to loss of funding are likewise an ongoing concern.<sup>15</sup>

# Alignment with Existing Roles and Capabilities

In addition to exploring the overall suitability of IRs to mirror DR collections, we further consider how the proposed service model aligns with data management roles and activities among libraries and librarians. Alignment is here considered from multiple perspectives, including administrative-level collaborations, the participation of functional and subject area librarians, and system capabilities.

With regard to collaboration, the development of sustainable data services can benefit from the engagement of library administration with stakeholders from their respective campus IT units and sponsored research offices. As a notable example, Witt describes the development of the Purdue University Research Repository (PURR),<sup>16</sup> an effort which was steered by a working group whose members included, among others, the Associate Vice President for Research, two Associate Deans from the Libraries, liaison librarians, and technical specialists.<sup>17</sup> Similarly composed groupings are proposed by Block and colleagues at Cornell University,<sup>18</sup> and the 2012 ACRL study by Tenopir and colleagues likewise highlights the experience among library directors that sponsored research units in particular are necessary contributors to the development of impactful RDM services.<sup>19</sup>

By collaborating at administrative levels to strategically position data and repository services within the research practice of an institution, the identification and promotion of the IR as a complementary service to DRs can become part of the research planning strategy. For example, Choudhury notes a particularly promising outcome of engaging university administrators in the development of the Johns Hopkins University Data Management Services (JHUDMS).<sup>20</sup> As a demonstration of the anticipated value the service may provide to researchers, the JHU administration opted to directly fund preproposal consultations between JHUDMS and researchers applying for grants. This consultation includes a review of domain repository options together with information about the JHU Data Archive.<sup>21</sup> Optional, grant-funded post-award services are also available that can include eventual transfer of data to the archive. This and similar arrangements are of particular import to the service proposed here as they logically extend to proactively defining complementary roles between DRs and IRs. By thoroughly reviewing repository options with researchers and mapping repository capabilities and features to different phases of the data life cycle, librarians are positioned to make strategic recommendations about when and under what circumstances the IR represents a viable option for data archiving. Although such consultations represent librarian rather than administrator activity, the sponsorship of the JHUDMS by university administration in this case demonstrates how a successful collaboration can lead to better promotion of the IR and facilitate collection development.

At the grassroots level, discussions of librarian roles in support of data curation may distinguish between subject area and functional expertise.<sup>22</sup> While both contexts may overlap within particular positions, a pairing or collaboration between subject and functional specialists as described by Jaguszewski and Williams is a promising strategy for providing both the domain and technical expertise to effectively support researchers.<sup>23</sup> For example, the composition of data curation project teams at Purdue, as reported by Newton and colleagues, demonstrate a distribution of functional skills and subject area expertise across an organization.<sup>24</sup> Other models exist, but the overall implication for IR building in this context is the importance of linking tangible capabilities with researcher needs.

On the functional side, as described for example by Tenopir and colleagues, Sands and colleagues, and Lyon, services performed by IR managers and data curation librarians can include transforming proprietary files to open file formats, conducting file integrity and format validation routines, creating or transforming metadata, and packaging data for submission to the IR.<sup>25</sup> These processes and activities will necessarily be important components of an IR data mirroring service. However, as noted by Kim, there remains nonetheless a growing imperative for technical assistance and “a more proactive role in support of digital scholarship” that is relevant to extending IR services.<sup>26</sup> Because the proposed model is focused on the batch transfer and repository ingest of complete data set collections, it may be necessary to scale up workflows that are currently oriented toward the curation of single or small collections of data sets. At minimum, adapting workflows in this way will require some scripting capabilities and familiarity with application programming interfaces (APIs).

It has likewise been shown that IR managers and data curation librarians are not necessarily technicians and that the duties of librarians in these positions may focus on assisting researchers with the identification and implementation of best practices in content, data, and metadata management. Lyle and colleagues, for example, describe a series of collaborations between the Inter-university Consortium for Political and Social Research (ICPSR) and multiple IRs to curate and publish legacy datasets.<sup>27</sup> Noting at the outset that many IR managers have “limited experience dealing with quantitative or qualitative data,”<sup>28</sup> the authors proceed through a series of case studies that highlight the types of functional support IR managers may need in preparing data for archiving. However, in lieu of technical skills, the strengths in relationship and resource building that participating librarians brought to the case studies indicated that IR managers and data curation librarians are well-positioned to mediate between data owners and developers or technicians in support of collection-scale curation and archiving.<sup>29</sup>

Established data management activities of subject area librarians can be likewise aligned with the proposed IR data mirroring service. As reported by Antell and colleagues, data management skills practiced with some regularity among librarians include consultation about DRs as well as providing information about

data life cycle management and funder requirements.<sup>30</sup> Similar to the JHUDMS example above, such consultations provide an opportunity for librarians to identify publication and archiving requirements that a local IR may appropriately provide in the absence of, or in addition to, an established DR. Additionally, Newton and colleagues described the value of the domain expertise that subject librarians bring to the selection and appraisal of data sets for IR inclusion,<sup>31</sup> while discussion in Bracke further illustrated the application of domain knowledge to support data curation and metadata development.<sup>32</sup>

All of these activities are relevant to extending IR service models, as subject librarians are well-positioned to know which DRs their faculty utilize and the long-term preservation capabilities and funding prospects of those repositories. This awareness is essential to identifying published data sets that may benefit from mirroring within the IR, as well as identifying “value add” services that the IR can provide, like supplementary documentation, citation linking, or other services. Similarly, because the DR mirroring service is oriented toward the batch curation and archiving of collections rather than toward individual datasets, the expertise that subject librarians bring to smaller-scale appraisals may more broadly carry over to assessing the long-term value of DR collections based on uniqueness or impact.

A final area of interest with regard to aligning existing IR capabilities with the proposed service relates to technical infrastructure. As noted above, repository solutions with wide adoption among libraries are strongly oriented toward traditional scholarly document types such as preprints and ETDs, with out-of-the-box support for a limited metadata profile based on the simple or qualified Dublin Core schemas.<sup>33</sup> Nonetheless, as reported by Carlson and colleagues and Johnston,<sup>34</sup> workflows have been developed that support data curation and publication within common IR platforms including Digital Commons and DSpace.<sup>35</sup>

In many cases, a lack of data-ready features within IRs can result in a flattening of complex metadata and a format-agnostic presentation of data formats and file types. Even so, expressed priorities and concerns of researchers demonstrate that the publication, permanent identification, and preservation features common among IR platforms can contribute to their adoption as data repositories. For example, Cragin and colleagues and McLure and colleagues described the differing perceptions of researchers regarding concerns and expectations for sharing data and the corresponding service implications for repository builders.<sup>36</sup> Limitations aside, important service capabilities as identified by Cragin and colleagues are well-supported by IR platforms, including embargoes and specification of use requirements with preferred citations.<sup>37</sup> McLure and colleagues likewise documented researcher views on the potential benefit of IRs as locally managed dissemination and preservation platforms.<sup>38</sup> By identifying service requirements of researchers that map to the general purpose, discipline-agnostic nature of IRs, such findings suggest a selective use of IRs to mirror DR collections is a valid use case in alignment with researcher priorities. Taken together with the conceptual



rationale provided above, establishing IR mirrors of DR collections can be of particular benefit when the partnering DR or its data providers lose funding. Additionally, when storage limitations or competing priorities require DRs to concentrate resources around high-use data, mirroring or transferring less in-demand data to an IR offers a means to maintain access through a distribution of management and stewardship duties.

## Applications: Requirements and Example Use Cases

Based on the above discussion a case can be made that IRs are suitable platforms to serve as mirrors of DR published data collections. That said, it's important to reiterate that a mirroring service is likely to be practical only if implemented through batch workflows, the development of which will be dependent upon differing DR architectures. However, for the purpose of defining an extensible process model, the scenarios and strategies below are organized into three broadly defined phases: defining stakeholder interactions and requirements, harvesting and metadata processing, and content curation and packaging.

## Defining Stakeholder Interactions and Requirements

The first phase of a mirroring service to reflect the contexts of a DR in your IR involves defining the stakeholder interactions and baseline requirements for harvest and ingest procedures. Among other things, the IR or project manager must determine how to satisfy the use, access, and attribution requirements of stakeholders representing the source DR. Minimally, this involves securing permission to harvest and republish the data, either formally via a submission agreement or informally through e-mail or verbal agreement. Additionally, details about which data to transfer along with a proposed schedule should be documented with the necessary authorizations. This documentation is similar to using a submission agreement.

If the data to be mirrored are not subject to restrictions that would prevent mirroring, such formal agreements may not be necessary. However, IR managers should be sensitive to the potential for confusion among researchers who originally contributed their data to the DR. While communicating about the project directly with the researchers or contributors may not be feasible or practical, regular communication with key DR stakeholders about the project

time line and milestones can help prevent misunderstandings. For example, following a collection ingest, IR managers may want to promote the mirroring project via a press release or mass e-mail to their campus community. Such communications should be timed so that researchers who contributed data to the DR are well-informed before any broader announcements are made to potential users.

Regarding use and access permissions, DRs may explicitly include permission information within the corresponding item-level metadata, or else the IR must work to translate this information from implicit repository or collection-level policies. For example, in 2015 the University of New Mexico (UNM) Libraries collaborated with the Sevilleta Long Term Ecological Research (LTER) program to archive and mirror data sets previously published in the LTER Network Data Portal.<sup>39</sup> Establishing the authority of the libraries to republish the Sevilleta LTER data via the IR was a multistep process of exploring different strategies for incorporating the LTER data policy. Ultimately, boilerplate language was included as rights metadata within item records with a reference to the full policy online.<sup>40</sup> Preferred citations referencing the original LTER version of the data were also copied into item records within the IR. Throughout the process, librarians consulted with LTER stakeholders and developed test collections to model different ways of presenting the information.

Another example is provided by Geographic Storage and Retrieval Engine (GSToRE), maintained by the Earth Data Analysis Center (EDAC) of UNM.<sup>41</sup> The GSToRE data are collected from a variety of sources, and there are no overarching access or use policies. Item-level permissions vary, and many data sets are public domain with no access or use constraints, though a boilerplate liability disclaimer inserted by EDAC encourages the citation of data sources.<sup>42</sup> However, because the preservation model in GSToRE is centered on exporting archive-ready packages to external systems, such as IRs, by implication mirroring collections is an anticipated and generally approved use. Importantly, prior to a harvest and ingest, IR managers or data librarians may refer to GSToRE documentation of service-level and other agreements provided to data depositors. As above, this information can be used to develop boilerplate statements for inclusion within data set metadata for any items mirrored within an IR.

Once stakeholder roles and any conditions for access and use have been addressed, the harvest and ingest process can be further broken down into defining and fulfilling requirements around metadata and content modeling. These requirements will often amount to technical compromises negotiated between the IR and DR. Therefore, it is useful to have access to a development server for prototyping. Testing the ingest procedures within a development environment will additionally allow IR managers to assess what, if any, impact a batch data set ingest may have on IR storage capacity and performance.

# Harvesting and Metadata Processing

Common scenarios for metadata harvest include automated retrieval via an API or more manual processes using a web crawler such as Wget.<sup>43</sup> Between the two, APIs are the preferred means of access where available; DRs may publish custom APIs or make use of standard APIs including the Simple Web-service Offering Repository Deposit (SWORD) protocol.<sup>44</sup> Many repository architectures likewise support the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a metadata-specific API that facilitates discovery and federation.<sup>45</sup>

For example, a popular turnkey data repository application that makes use of SWORD as well as OAI-PMH is the Dataverse Network (DVN). While the provision and maintenance of a Dataverse may not be within a given institution's capabilities, the system's growing adoption<sup>46</sup> together with its open and interoperable design may result in an increasing use of existing Dataverse Networks by faculty and researchers external to the hosting institution. In such cases, a potential service model for IR managers would be the aggregation of researchers' externally published data. As a means of previewing or analyzing corresponding metadata ahead of transferring data sets from a Dataverse to an IR, the DVN OAI-PMH interface can be accessed via a web browser or scripted using Wget, cURL, or other HTTP interfaces.<sup>47</sup> In cases where it's preferable to mirror just the metadata and maintain the external DVN as the canonical source of data files, DSpace and other IR applications include OAI-PMH utilities that allow metadata from external repositories to be harvested and published in minutes.

Harvesting metadata via custom APIs will likely be more complex, as IR managers or data librarians will need to develop the necessary software or scripts. For example a set of Python scripts have been developed for a planned harvest of GSToRE data for archiving in UNM's IR that will access canonical DR metadata via JSON and XML through the specifically developed repository and metadata access API functions published by GSToRE.<sup>48</sup>

Once the metadata has been harvested, then mapping or cross-walking activities to align the DR-provided metadata with the IR metadata schema can take place. Some schemas will be easy to map and will include descriptors that are synonymous with the IR metadata such as the simple Dublin Core elements "title," "description," and "publisher." However, even when fields from the source schema map to identically named fields in the destination schema, some analysis is necessary to determine if the fields are used in the same way. Especially when the source metadata schema includes a deeply nested hierarchy, IR managers will need to determine how best to represent multiple source fields that map to a single destination field. For example, the Data Documentation Initiative (DDI) Codebook standard defines unique fields for topic classes, keywords, study concepts, and coverage.<sup>49</sup> Many of these might

be cross-walked to the Dublin Core “subject” field, but concatenation of multiple source metadata fields could be noisy and negatively impact web displays or usability. Decisions about metadata mapping will be decided by IR capabilities and the preferences of the DR stakeholders. In some cases, documentation of best practices and recommended cross-walks will be available. Specific to the example given here, a DDI-to-Dublin Core cross-walk is provided by the DDI Alliance.<sup>50</sup>

If the IR capabilities include metadata extension or customization, another strategy is to map the source metadata to an alternative schema or use multiple schemas. For example, DSpace versions 1.5 and above support the registration of multiple “flat” schemas, which enable IR managers to combine fields from different schemas when describing items.<sup>51</sup> In practice, while a complex schema such as the Ecological Metadata Language (EML) cannot be fully cross-walked to the DSpace Dublin Core profile,<sup>52</sup> the standard can be mapped to Darwin Core in a semantically meaningful way.<sup>53</sup> Without utilizing a nested hierarchy of “coverage” fields as in EML, a single qualified term set in Darwin Core nonetheless includes categories of domain-specific terms such as “GeologicalContext” and “Taxon.” In support of mirroring data from repositories that use EML metadata, extending the DSpace metadata registry to implement Darwin Core is a simple process of registering the namespace URI and adding desired fields (figure 7.1).

The screenshot shows the 'Metadata registry' page in the UNM LoboVault interface. The page header includes the UNM LoboVault logo and the user name 'Jonathan Wheeler'. The main content area is titled 'Metadata registry' and contains a table of metadata schemas. The table has three columns: ID, Namespace, and Name. The schema with ID 8 is highlighted with a red box. Below the table is a 'Delete schema' button and an 'Add a new schema' form with fields for Namespace and Name.

ID	Namespace	Name	
1	http://dublincore.org/documents/dcmi-terms/	dc	
<input type="checkbox"/>	2	http://elibrary.unm.edu/embargo-terms/	emb
<input type="checkbox"/>	4	http://www.ndltd.org/standards/metadata/etdms/1.0/xml.xsd	etdms
<input type="checkbox"/>	5	http://elibrary.unm.edu/data	data
<input type="checkbox"/>	6	http://purl.org/dc/terms/	dc/terms
<input type="checkbox"/>	7	http://dspace.org/eperson	eperson
<input type="checkbox"/>	8	http://rs.tdwg.org/dwc/terms/index.htm	dwc

Below the table, there is a 'Delete schema' button and an 'Add a new schema' section with the following fields:

**Add a new schema**

**Namespace: \***

Namespace should be an established URI location for the new schema.

**Name: \***

**FIGURE 7.1**

The DSpace administrator’s view of the metadata registry. The Darwin Core namespace is highlighted.

The benefits of this approach are demonstrated by a map visualization feature within UNM's IR that was developed in support of the Sevilleta LTER data-archiving project. Because of the important geographical context of the data, it was desirable to reproduce the maps drawn by the network portal for items with coordinate metadata. Using the qualified Dublin Core "spatial coverage" element was impractical because existing items already used that field to provide place names, and mixing coordinate and text data types would have broken the JavaScript/XSL mapping template developed for UNM's DSpace instance. By extending the metadata registry to include the Darwin Core "decimalLatitude" and "decimalLongitude" elements, librarians were able to enforce a coordinate data constraint within the mapping template (figure 7.2).

The screenshot shows the 'Metadata Schema: "dwc"' interface. On the left is a navigation sidebar with sections: UNM LIBRARIES (University Libraries, Law Library, Health Sciences Library), BROWSE (All of LoboVault, Communities & Collections, Date, Authors, Titles, Subjects), MY ACCOUNT (My Exports, Logout, Profile, Submissions), and ADMINISTRATIVE (Control Panel, Statistics, Curation Tasks, Access Control, People, Groups, Authorizations, Content Administration). The main content area is titled 'Metadata Schema: "dwc"' and contains a description: 'This is the metadata schema for "http://rs.tdwg.org/dwc/terms/index.htm". You may add new or update existing metadata fields to this schema. Fields may also be selected for deletion or be moved to another schema.' Below this is a section 'Add new metadata field' with a 'Field Name: dwc .' label and two empty input fields. A 'Scope Note:' label is followed by a large text area. Below the text area is a button 'Add new metadata field'. A section 'Schema metadata fields' contains a table with two rows:

ID	Field	Scope Note
<input type="checkbox"/> 160	dwc.decimalLatitude	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.
<input type="checkbox"/> 161	dwc.decimalLongitude	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.

Below the table are buttons: 'Delete fields', 'Move fields to another schema', and 'Return'.

**FIGURE 7.2**  
Adding fields for metadata schema.

Once decisions about representing DR metadata within the IR have been made, the harvested metadata content must be cross-walked to the IR schema and saved in a file format accepted by the IR for batch ingest. Typically, this pro-

cess will be accomplished using XSL templates to transform XML metadata, but other options may exist. DSpace, for example, allows batch creation and editing of metadata via CSV upload through the web interface.<sup>54</sup>

## Content Curation and Packaging

Finally, together with its metadata schema, the destination IR will have specific requirements for associating content files with their respective metadata and packaging items for ingest. As with metadata, content files can be harvested through a variety of means, preferably via API but alternatively through batch HTTP requests via `cURL` or `Wget`. Whichever method is used, an important pre-harvest activity is to create an inventory of the DR assets to be acquired. This information, which may be published as site statistics or requested from DR administrators, minimally provides a quick overview of item and version counts that can be used later to verify the completeness of the harvest. In addition to an inventory, librarians managing a harvest must also identify the file validation scheme used within the DR. For example, checksums will often be made accessible via API and should be used to validate harvested files.

Wherever possible, IR managers and librarians should seek to curate the data for preservation and explore options for otherwise adding value to the data and metadata. Minimally, curation will involve documenting and exposing provenance information relevant to the mirroring process, such as the date of harvest and the outcomes of virus scans, file validation, and format identification routines. These processes may be readily incorporated into a collection-scale workflow through the use of existing batch utilities like the Digital Record Object Identification (DROID) software tool.<sup>55</sup>

Further curation actions may include compiling any additional documentation necessary to support data discovery and use within the IR context. For example, an early and relatively small batch data ingest into the UNM IR involved mirroring a set of colonia population data published by the Bureau of Business and Economic Research (BBER).<sup>56</sup> In communication with the lead researcher, the content files were harvested from the BBER website using `Wget`, and the metadata and supporting documentation were compiled through discussion and by collating any corresponding presentations, reports, and so on. Additional curation activities performed on the data set included transforming files from proprietary formats to open formats and creating provenance and technical metadata using DROID and a locally developed METS utility.<sup>57</sup> These and other value-add activities resulted in the publication of an IR mirror of the BBER data set that was more than just a duplication of the original resource.<sup>58</sup> Also, the simple but scalable batch workflow was a prototype of the procedures used to curate and package the Sevilleta LTER data.

For the final ingest into the IR, item- and collection-level packaging requirements will be platform-dependent. Consequently, the role or involvement of the repository manager will vary according to whether the IR is hosted by a third party, locally maintained, or open source. While the IR manager's participation in batch ingest routines within proprietary systems may be limited, the necessary features should exist, and vendors are often interested in exploring innovative uses of their systems. As an example, Carlson and colleagues reported on a project in which materials from a large research center were curated within a bepress Digital Commons repository at Purdue.<sup>59</sup>

Alternatively, managers of locally hosted, open-source platforms such as DSpace may capitalize on available documentation and utilities developed by the user community. Specifically, DSpace supports batch ingest of items packaged according to a Simple Archive Format (SAF) specification.<sup>60</sup> Similar to the Bagit digital content transfer utility developed by the Library of Congress,<sup>61</sup> SAF describes a per-item file structure and automated ingest process for DSpace repositories. The available documentation is comprehensive, but in summary a collection packaged for ingest using SAF will consist of a directory or zip archive containing individual, item-level subdirectories. The subdirectories will contain the item's associated content files, one or more XML metadata files, a text file manifest describing the content file types, and, optionally, a text file designating the collection or collections to which the item belongs. Ingest is completed by submitting SAF packages to the repository via a command line utility or, alternatively, using a web-based batch import feature introduced in DSpace version 5.<sup>62</sup>

Following ingest, some post-processing for quality assurance purposes is recommended. In addition to verifying that the process concluded without errors, quality checks can range in scope and depth and can be implemented through various manual or automated processes. For example, following ingest of the BBER colonia data, the relatively small size of the data set enabled librarians to perform manual quality checks. These checks included downloading the individual files to identify file formats and validate checksums using a second run through DROID. In the case of the LTER data ingest, a percentage of the collection was manually checked for format and checksum validation, but automated processes were run against the full collection using available DSpace curation tools. These tools include file format identification and checksum validation processes that may be run on demand against an item, collection, or community. None of the quality checks performed on either the BBER or LTER collections identified any errors. However, because batch processing can result in the propagation of errors across an entire collection, such follow-up checks are an important element of a harvest and ingest workflow.



# Conclusion

As researcher and institutional data management needs evolve to encompass federal public access planning and DMP compliance requirements, the demand for library data management services may be expected to grow accordingly. In addition to well-established activities such as DMP consultation and data reference, technical support for asset management and data preservation represent additional niche services that academic libraries are well-situated to provide based on existing professional skill sets, established IR infrastructures, and corresponding digital preservation workflows. While near-term sharing and timely publication of data via DRs remains a researcher-preferred strategy, the migration or mirroring of previously published data within IRs may provide capabilities in support of archiving and reuse that are complementary or supplementary to DR publication features. Although such mirroring represents a promising service model for libraries, the potential for incorporating a routine collection-scale ingest activity requires the corresponding development of batch harvest, packaging, and ingest processes.

While acknowledging that the workflows presented here are desktop-based and thus do not fully address scalability issues, there are some advantages to maintaining desktop workflows, such as quality control. Further, the curation and packaging of collections is similar to curating individual data sets in that it is a high-touch activity and requirements will vary from case to case. Because of this, the need to customize processes will inevitably impact scalability. However, bandwidth issues and storage constraints will present themselves, and a future development of flexible utilities for data transfer between DRs and IRs is needed. In particular, as initiatives such as the Digital Preservation Network (DPN) grow,<sup>63</sup> a near-term focus for IR managers should be the development of processes that automatically generate archival information packages for DR data on harvest. Because not all IRs are maintained as archival or preservation platforms, such a feature would enable a parallel transfer of DR data collections to alternative preservation services such as DPN and DuraCloud.<sup>64</sup> Through development of these and other services to better position IRs within the web of repositories, the collective contribution of libraries to data preservation will further demonstrate their value as memory institutions and partners within a global data infrastructure.

# Acknowledgments

The author would like to thank and acknowledge the following for their feedback and contributions during the Sevilleta LTER data ingest into UNM's institutional repository: Kristin Vanderbilt (Sevilleta LTER Program), Mark Servilla (LTER Network Office), Jacob Nash (UNM Health Sciences Library and Informatics Center), and UNM Libraries Information Technology Services.

The Python and XSLT scripts used to harvest and package the Sevilleta LTER data for ingest into a DSpace repository are available at <https://lobogit.unm.edu/jwheel01/lter-collection-harvest>.

## Notes

1. Karen Antell, Jody Bales Foote, Jaymie Turner, and Brian Shults, "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions," *College and Research Libraries* 75, no. 4 (July 2014): 557–74, doi:10.5860/crl.75.4.557. Regarding planning, implementation, and perceived value of RDS services among ACRL libraries, an interesting corollary discussion of how perceptions align between library administrators and librarians is provided in Carol Tenopir, Robert J. Sandusky, Suzie Allard, and Ben Birch, "Research Data Management Services in Academic Research Libraries and Perceptions of Librarians," *Library and Information Science Research* 36, no. 2 (April 2014): 84–90, doi:10.1016/j.lisr.2013.11.003.
2. Antell et al., "Dealing with Data"; Hans Jørn Nielsen and Birger Hjørland, "Curating Research Data: The Potential Roles of Libraries and Information Professionals," *Journal of Documentation* 70, no. 2 (2014): 221–240, doi:10.1108/JD-03-2013-0034.
3. Don MacMillan, "Data Sharing and Discovery: What Librarians Need to Know," *Journal of Academic Librarianship* 40, no. 5 (September 2014): 546, doi:10.1016/j.acalib.2014.06.011.
4. Nancy Y. McGovern and Aprille C. McKay, "Leveraging Short-Term Opportunities to Address Long-Term Obligations: A Perspective on Institutional Repositories and Digital Preservation Programs," *Library Trends* 57, no. 2 (2008): 262–79, <https://muse.jhu.edu/article/262030>; Priti Jain, "New Trends and Future Applications/Directions of Institutional Repositories in Academic Institutions," *Library Review* 60, no. 2 (March 2011): 125–41, doi:10.1108/00242531111113078.
5. Karen S. Baker and Lynn Yarmey, "Data Stewardship: Environmental Data Curation and a Web-of-Repositories," *International Journal of Digital Curation* 4, no. 2 (October 15, 2009): 12–27, doi:10.2218/ijdc.v4i2.90; Karen S. Baker and Florence Millerand, "Infrastructuring Ecology: Challenges in Achieving Data Sharing," in *Collaboration in the New Life Sciences*, ed. John N. Parker, Niki Vermeulen, and Bart Penders (Burlington, VT: Ashgate, 2010), 111–38.
6. Tyler Walters, "Assimilating Digital Repositories into the Active Research Process," in *Research Data Management: Practical Strategies for Information Professionals*, ed. Joyce M. Ray (West Lafayette, IN: Purdue University Press, 2014), eBook Collection, EBSCOhost, ISBN 9781461956815. Accessed February 19, 2016.
7. Greg Janée, Justin Mathena, and James Frew, "A Data Model and Architecture for Long-Term Preservation," in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: ACM, 2008), 134–44. doi:10.1145/1378889.1378912.
8. National Science Foundation, Grant Proposal Guide, Chapter II, Proposal Preparation Instructions, Section C.2.j, last modified January 25, 2016, accessed March 23, 2016, [http://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg\\_2.jsp#IIC2j](http://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_2.jsp#IIC2j). The DMP requirement described within the Grant Proposal Guide (GPG) includes separate recommendations covering "policies for access and sharing" and "plans for archiving data, samples, and other research products, and for preservation of access to them"; US

- Department of Energy, Office of Science “Statement on Digital Data Management,” last modified July 28, 2014, accessed March 22, 2016, <http://science.energy.gov/funding-opportunities/digital-data-management>.
9. G. Sayeed Choudhury, “Case Study 1: Johns Hopkins University Data Management Services,” in *Delivering Research Data Management Services*, ed. Graham Pryor, Sarah Jones, and Angus Whyte (London: Facet Publishing, 2014), 118.
  10. Paul F. Uhlir, “Information Gulags, Intellectual Straightjackets, and Memory Holes: Three Principles to Guide the Preservation of Scientific Data,” *Data Science Journal* 9 (2010): ES5. [https://www.jstage.jst.go.jp/article/dsj/9/0/9\\_Essay-001-Uhlir/\\_article](https://www.jstage.jst.go.jp/article/dsj/9/0/9_Essay-001-Uhlir/_article).
  11. Key Perspectives Ltd., *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability. SCARP Synthesis Study* (Digital Curation Centre, 2010), <http://hdl.handle.net/1842/3364>.
  12. Donatella Castelli, Paolo Manghi, and Costantino Thanos, “A Vision towards Scientific Communication Infrastructures: On Bridging the Realms of Research Digital Libraries and Scientific Data Centers,” *International Journal on Digital Libraries* 13, no. 3–4 (September 2013): 155–69, doi:10.1007/s00799-013-0106-7.
  13. *Ibid.*, 162.
  14. *Ibid.*, 162–163.
  15. Key Perspectives, *Data Dimensions*.
  16. Michael Witt, “Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service,” *Journal of Library Administration* 52, no. 2 (2012): 172–88; Purdue University Research Repository, accessed March 23, 2016, <https://purr.purdue.edu/>.
  17. Witt, “Co-designing,” 176.
  18. William C. Block, Eric Chen, Jim Cordes, Dianne Dietrich, Dean B Krafft, Stefan Kramer, David Lifka, Janet McCue, and Gail Steinhart, *Meeting Funders’ Data Policies: Blueprint for a Research Data Management Service Group (RDMSG)*, project report (Ithaca, NY: Cornell University, 2010), <http://hdl.handle.net/1813/28570>.
  19. Carol Tenopir, Ben Birch, and Suzie Allard, *Academic Libraries and Research Data Services*, an ACRL white paper (Chicago: Association of College and Research Libraries, 2012), 37–39.
  20. Choudhury, “Case Study 1,” 128.
  21. Johns Hopkins Data Archive Dataverse Network, accessed March 23, 2016, <https://archive.data.jhu.edu/dvn/>.
  22. Tenopir et al., “Research Data Management,” 87, provides an example of a distinction between informational (or consulting) RDS and technical RDS.
  23. Janice Jaguszewski and Karen Williams, *New Roles for New Times*, report (Washington, DC: Association of Research Libraries, August 2013), 13, <http://hdl.handle.net/11299/169867>.
  24. Mark P. Newton, C. C. Miller, and Marianne Stowell Bracke, “Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force,” *Collection Management* 36, no. 1 (2010): 53–67, doi:10.1080/01462679.2011.530546.
  25. Tenopir et al, “Research Data Management,” 87; Ashley E. Sands, Christine L. Borgman, Sharon Traweek, and Laura A. Wynholds, “We’re Working on It: Transferring the Sloan Digital Sky Survey from Laboratory to Library,” *International Journal of Digital Curation* 9, no. 2 (October 30, 2014), doi:10.2218/ijdc.v9i2.336; Liz Lyon, “The Informatics Transform: Re-engineering Libraries for the Data Decade,” *International Journal of Digital Curation* 7, no. 1 (March 12, 2012): 131–32, doi:10.2218/ijdc.v7i1.220.

26. Jeonghyun Kim, "Data Sharing and Its Implications for Academic Libraries," *New Library World* 114, no. 11/12 (November 18, 2013): 503, doi:10.1108/NLW-06-2013-0051.
27. Jared Lyle, George Alter, and Ann Green, "Partnering to Curate and Archive Social Science Data," in *Research Data Management: Practical Strategies for Information Professionals*, ed. Joyce M. Ray (West Lafayette, IN: Purdue University Press, 2014) eBook Collection, EBSCOhost, ISBN 9781461956815 accessed February 19, 2016; Inter-university Consortium for Political and Social Research homepage, accessed March 23, 2016, <https://www.icpsr.umich.edu/icpsrweb/landing.jsp>.
28. Lyle, Alter, and Green, "Partnering to Curate and Archive Social Science Data," under the heading "Need for Support."
29. Ibid., under the heading "Find."
30. Antell et al., "Dealing with Data," 567.
31. Newton, Miller, and Bracke, "Librarian Roles," 58, 61.
32. Marianne Stowell Bracke, "Emerging Data Curation Roles for Librarians: A Case Study of Agricultural Data," *Journal of Agricultural and Food Information* 12, no. 1 (2011): 65–74, doi:10.1080/10496505.2011.539158.
33. For example information about DSpace's default metadata schema see DSpace, "Functional Overview," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Functional+Overview#FunctionalOverview-MetadataManagement>. Similar information for bepress's Digital Commons is available at bepress, "Metadata Options in Digital Commons," Digital Commons Reference Material and User Guides, last modified January 2016, accessed March 23, 2016, <http://digitalcommons.bepress.com/cgi/viewcontent.cgi?article=1095&context=reference>.
34. Jake Carlson, Alexis E. Ramsey, and J. David Kotterman, "Using an Institutional Repository to Address Local-scale Needs: A Case Study at Purdue University," *Library Hi Tech* 28, no. 1 (March 9, 2010): 152–73, doi:10.1108/07378831011026751; Lisa R. Johnston, *A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot* (University of Minnesota Digital Conservancy, 2014), <http://hdl.handle.net/11299/162338>.
35. Digital Commons homepage, accessed March 23, 2016, <http://digitalcommons.bepress.com/>; DSpace homepage, accessed March 23, 2016, <http://dspace.org/>.
36. Melissa H. Cragin, Carole L. Palmer, Jacob R. Carlson, and Michael Witt, "Data Sharing, Small Science and Institutional Repositories," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, no. 1926 (September 13, 2010): 4023–38, doi:10.1098/rsta.2010.0165; Merinda McLure, Allison V. Level, Catherine L. Cranston, Beth Oehlerts, and Mike Culbertson, "Data Curation: A Study of Researcher Practices and Needs," *portal: Libraries and the Academy* 14, no. 2 (2014): 139–64, doi:10.1353/pla.2014.0009.
37. Cragin et al., "Data Sharing," 4035–36.
38. McLure et al., "Data Curation," 154.
39. Sevilleta Long Term Ecological Research Program, accessed March 23, 2016, <http://sev.lternet.edu/>; LTER Network Data Portal, accessed March 23, 2016, <https://portal.lternet.edu/nis/home.jsp>.
40. Long Term Ecological Research Network, "LTER Network Data Access Policy, Data Access Requirements, and General Data Use Agreement," accessed March 23, 2016, <http://www.lternet.edu/policies/data-access>.

41. GStoRE (Geographic Storage, Transformation and Retrieval Engine), version 3, homepage, accessed March 23, 2016, <https://gstore.unm.edu/>; EDAC (Earth Data Analysis Center) homepage, accessed March 23, 2016, <http://edac.unm.edu/>.
42. See, for example, the section on “Distribution Liability” in GStoRE “Wildfire Risk Main Model,” accessed March 22, 2016, <http://gstore.unm.edu/apps/rgis/datasets/71be383b-ad19-4252-9c01-cfad3216a0ca/metadata/FGDC-STD-001-1998.html>.
43. “GNU Wget 1.18 Manual,” last modified December 11, 2015, accessed March 23, 2016, <https://www.gnu.org/software/wget/>.
44. SWORD homepage, accessed March 23, 2016, <http://swordapp.org/>.
45. Open Archives Initiative, “Protocol for Metadata Harvesting,” accessed March 23, 2016, <https://www.openarchives.org/pmh/>.
46. Dataverse Project, “Dataverse Repositories,” accessed March 22, 2016, <http://dataverse.org/>. This overview on the Dataverse Project website shows fourteen Dataverse repositories worldwide as of March 11, 2016. It should be noted, however, that an individual repository may host Dataverses for other institutions. For example, the Harvard repository includes over 1,400 “sub” Dataverses, many of which are sponsored by external universities and organizations. In Europe, the Utrecht University’s DataverseNL likewise hosts Dataverses sponsored by institutions throughout Central and Eastern Europe.
47. cURL homepage, accessed March 23, 2016, <https://curl.haxx.se/>.
48. python homepage, accessed March 23, 2016, <https://www.python.org/>.
49. Data Documentation Initiative, “DDI-Codebook 2.5,” accessed March 23, 2016, <http://www.ddialliance.org/Specification/DDI-Codebook/2.5/>.
50. Data Documentation Initiative, “Mapping to Dublin Core (DDI Version 2),” accessed March 23, 2016, <http://www.ddialliance.org/resources/ddi-profiles/dc>.
51. Available since at least version 1.5, documentation for DSpace version 5 is available at DSpace, “Functional Overview,” DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Functional+Overview#FunctionalOverview-MetadataManagement>.
52. The Knowledge Network for Biocomplexity, accessed March 24, 2016, <https://knb.ecoinformatics.org/>.
53. Darwin Core Task Group, “Darwin Core,” issued February 12, 2009, last updated June 5, 2015, accessed March 23, 2016, <http://rs.tdwg.org/dwcl/>.
54. Available since DSpace version 1.6, documentation for the current version is available at DSpace, “Batch Metadata Editing,” DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Batch+Metadata+Editing>.
55. National Archives, “File Profiling Tool (DROID),” accessed March 23, 2016, <http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>.
56. Bureau of Business and Economic Research homepage, accessed March 23, 2016, <http://bber.unm.edu/>.
57. Library of Congress, “METS: Metadata Encoding and Transmission Standard,” last modified February 9, 2016, accessed March 23, 2016, <http://www.loc.gov/standards/mets/>.
58. Daren Ruiz, “Colonia Population and Socioeconomic and Housing Characteristic Estimates, Maps and Shape File Update: November 2012 [data set],” University of New Mexico (2012), <http://hdl.handle.net/1928/22547>. Additional content and metadata available at <http://repository.unm.edu/archive/Projects/22547/>.

59. Carlson, Ramsey, and Kotterman, "Using an Institutional Repository."
60. DSpace, "Importing and Exporting Items via Simple Archive Format," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format>.
61. Library of Congress, "Bagit: Transferring Content for Digital Preservation," video, 3:14, posted June 24, 2009, accessed March 23, 2016, <http://www.digitalpreservation.gov/multimedia/videos/bagit0609.html>.
62. DSpace, "Latest Release," accessed March 23, 2016, <http://www.dspace.org/latest-release>.
63. Digital Preservation Network homepage, accessed March 23, 2016, <http://dpn.org/>.
64. Duracloud homepage, accessed March 23, 2016, <http://www.duracloud.org/>.

## Bibliography

- Antell, Karen, Jody Bales Foote, Jaymie Turner, and Brian Shults. "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions." *College and Research Libraries* 75, no. 4 (July 2014): 557–74. doi:10.5860/crl.75.4.557.
- Baker, Karen S., and Florence Millerand. "Infrastructuring Ecology: Challenges in Achieving Data Sharing." In *Collaboration in the New Life Sciences*. Edited by John N. Parker, Niki Vermeulen, and Bart Penders, 111–38. Burlington, VT: Ashgate, 2010.
- Baker, Karen S., and Lynn Yarmey. "Data Stewardship: Environmental Data Curation and a Web-of-Repositories." *International Journal of Digital Curation* 4, no. 2 (October 15, 2009): 12–27. doi:10.2218/ijdc.v4i2.90.
- bepress, "Metadata Options in Digital Commons," Digital Commons Reference Material and User Guides, last modified January 2016, accessed March 23, 2016, <http://digitalcommons.bepress.com/cgi/viewcontent.cgi?article=1095&context=reference>.
- Block, William C., Eric Chen, Jim Cordes, Dianne Dietrich, Dean B. Krafft, Stefan Kramer, David Lifka, Janet McCue, and Gail Steinhart. *Meeting Funders' Data Policies: Blueprint for a Research Data Management Service Group (RDMSG)*. Project report. Ithaca, NY: Cornell University, 2010. <http://hdl.handle.net/1813/28570>.
- Bracke, Marianne Stowell. "Emerging Data Curation Roles for Librarians: A Case Study of Agricultural Data." *Journal of Agricultural and Food Information* 12, no. 1 (2011): 65–74. doi:10.1080/10496505.2011.539158.
- Bureau of Business and Economic Research, accessed March 23, 2016, <http://bber.unm.edu/>.
- Carlson, Jake, Alexis E. Ramsey, and J. David Kotterman. "Using an Institutional Repository to Address Local-Scale Needs: A Case Study at Purdue University." *Library Hi Tech* 28, no. 1 (March 9, 2010): 152–73. doi:10.1108/07378831011026751.
- Castelli, Donatella, Paolo Manghi, and Costantino Thanos. "A Vision towards Scientific Communication Infrastructures: On Bridging the Realms of Research Digital Libraries and Scientific Data Centers." *International Journal on Digital Libraries* 13, no. 3–4 (September 2013): 155–69. doi:10.1007/s00799-013-0106-7.
- Choudhury, G. Sayeed. "Case Study 1: Johns Hopkins University Data Management Services." In *Delivering Research Data Management Services*. Edited by Graham Pryor, Sarah Jones, and Angus Whyte, 115–33. London: Facet Publishing, 2014.

- Cragin, Melissa H., Carole L. Palmer, Jacob R. Carlson, and Michael Witt. "Data Sharing, Small Science and Institutional Repositories." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, no. 1926 (September 13, 2010): 4023–38. doi:10.1098/rsta.2010.0165.
- cURL, accessed March 23, 2016, <https://curl.haxx.se/>.
- Darwin Core Task Group, "Darwin Core," issued February 12, 2009, last updated June 5, 2015, accessed March 23, 2016, <http://rs.tdwg.org/dwc/>.
- Data Documentation Initiative, "DDI-Codebook 2.5," accessed March 23, 2016, <http://www.ddialliance.org/Specification/DDI-Codebook/2.5/>.
- Data Documentation Initiative, "Mapping to Dublin Core (DDI Version 2)," accessed March 23, 2016, <http://www.ddialliance.org/resources/ddi-profiles/dc>.
- Dataverse Project, "Dataverse Repositories," accessed March 22, 2016, <http://dataverse.org/>.
- Digital Commons, accessed March 23, 2016, <http://digitalcommons.bepress.com/>.
- Digital Preservation Network, accessed March 23, 2016, <http://dpn.org/>.
- DSpace, accessed March 23, 2016, <http://dspace.org/>.
- DSpace, "Batch Metadata Editing," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Batch+Metadata+Editing>.
- DSpace, "Functional Overview," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Functional+Overview#Functional+Overview-Metadamanagement>.
- DSpace, "Importing and Exporting Items via Simple Archive Format," DSpace 5.x Documentation, accessed March 23, 2016, <https://wiki.duraspace.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format>.
- DSpace, "Latest Release," accessed March 23, 2016, <http://www.dspace.org/latest-release>.
- Duracloud, accessed March 23, 2016, <http://www.duracloud.org/>.
- EDAC (Earth Data Analysis Center) homepage, accessed March 23, 2016, <http://edac.unm.edu/>.
- GNU Wget 1.18 Manual," last modified December 11, 2015, accessed March 23, 2016, <https://www.gnu.org/software/wget/>.
- GSToRE (Geographic Storage, Transformation and Retrieval Engine), version 3, homepage, accessed March 23, 2016, <https://gstore.unm.edu/>.
- GSToRE "Wildfire Risk Main Model," accessed March 22, 2016, <http://gstore.unm.edu/apps/rgis/datasets/71be383b-ad19-4252-9c01-cfad3216a0ca/metadata/FG-DC-STD-001-1998.html>.
- Inter-university Consortium for Political and Social Research, accessed March 23, 2016, <https://www.icpsr.umich.edu/icpsrweb/landing.jsp>.
- Jaguszewski, Janice, and Karen Williams. *New Roles for New Times: Transforming Liaison Roles in Research Libraries*. Report. Washington, DC: Association of Research Libraries, August 2013. <http://hdl.handle.net/11299/169867>.
- Jain, Priti. "New Trends and Future Applications/Directions of Institutional Repositories in Academic Institutions." *Library Review* 60, no. 2 (March 2011): 125–41. doi:10.1108/00242531111113078.
- Janée, Greg, Justin Mathena, and James Frew. "A Data Model and Architecture for Long-Term Preservation." In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 134–44. New York: ACM, 2008. doi:10.1145/1378889.1378912.
- Johns Hopkins Data Archive Dataverse Network, accessed March 23, 2016, <https://archive.data.jhu.edu/dvn/>.



- Johnston, Lisa R. *A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot*. University of Minnesota Digital Conservancy, 2014. <http://hdl.handle.net/11299/162338>.
- Key Perspectives Ltd. *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long Term Viability. SCARP Synthesis Study*. Digital Curation Centre, 2010. <http://hdl.handle.net/1842/3364>.
- Kim, Jeonghyun. "Data Sharing and Its Implications for Academic Libraries." *New Library World* 114, no. 11/12 (November 18, 2013): 494–506. doi:10.1108/NLW-06-2013-0051.
- Knowledge Network for Biocomplexity, accessed March 24, 2016, <https://knb.ecoinformatics.org/>.
- Library of Congress, "Bagit: Transferring Content for Digital Preservation," video, 3:14, posted June 24, 2009, accessed March 23, 2016, <http://www.digitalpreservation.gov/multimedia/videos/bagit0609.html>.
- Library of Congress, "METS: Metadata Encoding and Transmission Standard," last modified February 9, 2016, accessed March 23, 2016, <http://www.loc.gov/standards/mets/>.
- Long Term Ecological Research Network, "LTER Network Data Access Policy, Data Access Requirements, and General Data Use Agreement," accessed March 23, 2016, <http://www.lternet.edu/policies/data-access>.
- LTER Network Data Portal, accessed March 23, 2016, <https://portal.lternet.edu/nis/home.jsp>.
- Lyle, Jared, George Alter, and Ann Green. "Partnering to Curate and Archive Social Science Data." in *Research Data Management: Practical Strategies for Information Professionals*. Edited by Joyce M. Ray. West Lafayette, IN: Purdue University Press, 2014. eBook Collection, EBSCOhost, ISBN 9781461956815. Accessed February 19, 2016.
- Lyon, Liz. "The Informatics Transform: Re-engineering Libraries for the Data Decade." *International Journal of Digital Curation* 7, no. 1 (March 12, 2012): 126–38. doi:10.2218/ijdc.v7i1.220.
- MacMillan, Don. "Data Sharing and Discovery: What Librarians Need to Know." *Journal of Academic Librarianship* 40, no. 5 (September 2014): 541–49. doi:10.1016/j.acalib.2014.06.011.
- McGovern, Nancy Y., and Aprille C. McKay. "Leveraging Short-Term Opportunities to Address Long-Term Obligations: A Perspective on Institutional Repositories and Digital Preservation Programs." *Library Trends* 57, no. 2 (2008): 262–79. <https://muse.jhu.edu/article/262030>.
- McLure, Merinda, Allison V. Level, Catherine L. Cranston, Beth Oehlerts, and Mike Culbertson. "Data Curation: A Study of Researcher Practices and Needs." *portal: Libraries and the Academy* 14, no. 2 (2014): 139–64. doi:10.1353/pla.2014.0009.
- National Archives, "File Profiling Tool (DROID)," accessed March 23, 2016, <http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/>.
- National Science Foundation, Grant Proposal Guide, Chapter II, Proposal Preparation Instructions, Section C.2.j, last modified January 25, 2016, accessed March 23, 2016, [http://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg\\_2.jsp#IIC2j](http://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_2.jsp#IIC2j).
- Newton, Mark P., C. C. Miller, and Marianne Stowell Bracke. "Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force." *Collection Management* 36, no. 1 (2010): 53–67. doi:10.1080/01462679.2011.530546.

- Nielsen, Hans Jørn, and Birger Hjørland, "Curating Research Data: The Potential Roles of Libraries and Information Professionals," *Journal of Documentation* 70, no. 2 (2014): 221–40, doi:10.1108/JD-03-2013-0034.
- Open Archives Initiative, "Protocol for Metadata Harvesting," accessed March 23, 2016, <https://www.openarchives.org/pmh/>.
- Purdue University Research Repository, accessed March 23, 2016, <https://purr.purdue.edu/python>, accessed March 23, 2016, <https://www.python.org/>.
- Ruiz, Daren. "Colonia Population and Socioeconomic and Housing Characteristic Estimates, Maps and Shape File Update: November 2012 [data set]." University of New Mexico (2012). <http://hdl.handle.net/1928/22547>.
- Sands, Ashley E., Christine L. Borgman, Sharon Traweek, and Laura A. Wynholds. "We're Working on It: Transferring the Sloan Digital Sky Survey from Laboratory to Library." *International Journal of Digital Curation* 9, no. 2 (October 30, 2014). doi:10.2218/ijdc.v9i2.336.
- Sevilleta Long Term Ecological Research Program, accessed March 23, 2016, <http://sevilleta.net/>.
- SWORD, accessed March 23, 2016, <http://swordapp.org/>.
- Tenopir, Carol, Ben Birch, and Suzie Allard. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*. An ACRL white paper. Chicago: Association of College and Research Libraries, 2012.
- Tenopir, Carol, Robert J. Sandusky, Suzie Allard, and Ben Birch. "Research Data Management Services in Academic Research Libraries and Perceptions of Librarians." *Library and Information Science Research* 36, no. 2 (April 2014): 84–90. doi:10.1016/j.lisr.2013.11.003.
- Uhlir, Paul F. "Information Gulags, Intellectual Straightjackets, and Memory Holes: Three Principles to Guide the Preservation of Scientific Data" *Data Science Journal* 9 (2010): ES1–5. [https://www.jstage.jst.go.jp/article/dsj/9/0/9\\_Essay-001-Uhlir/\\_article.US](https://www.jstage.jst.go.jp/article/dsj/9/0/9_Essay-001-Uhlir/_article.US) Department of Energy, Office of Science. "Statement on Digital Data Management." Last modified July 28, 2014, accessed March 22, 2016. <http://science.energy.gov/funding-opportunities/digital-data-management>.
- Walters, Tyler. "Assimilating Digital Repositories into the Active Research Process." In *Research Data Management: Practical Strategies for Information Professionals*. Edited by Joyce M. Ray. West Lafayette, IN: Purdue University Press, 2014. eBook Collection, EBSCOhost, ISBN 9781461956815. Accessed February 19, 2016.
- Witt, Michael. "Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service." *Journal of Library Administration* 52, no. 2 (2012): 172–88.



## CHAPTER 8

# Beyond Cost Recovery

## Revenue Models and Practices for Data Repositories in Academia

*Karl Nilsen*

### Introduction

The economic sustainability of research data repositories is a complex problem for academic libraries that involves cost management and revenue generation. Yet the volume of published research and examples about revenue models for data repositories is considerably smaller than that on cost models. Revenue models address the sources of income for a data repository, whereas cost models typically provide a framework for describing, analyzing, and predicting the expenditures—such as technology and labor—associated with running a repository. Data repository managers can take advantage of a variety of cost frameworks,<sup>1</sup> consult a sizable body of literature on the costs of curation and preservation,<sup>2</sup> and, increasingly, review actual expenditures at other repositories.<sup>3</sup> In contrast, there are few guides to conventional and novel sources of income for data repositories. Fortunately, among the resources available are surveys by Maron, Kitchin and colleagues, Wang and colleagues, and Erway and Rinehart that shed light on revenue strategies for digital curation organizations.<sup>4</sup> This chapter builds on those contributions to enlarge our picture of revenue models for library-based data repositories and stimulate more discussion and debate about applicable business models. I use literature and public information to examine actual revenue practices at several data repositories and consider the advantages and disadvantages of each model

from the perspectives of repository managers and users. While I concentrate on library-based data repositories, the revenue practices at these institutions are not especially diverse, so I occasionally draw on examples from domain-specific repositories, such as Dryad and the Inter-university Consortium for Political and Social Research (ICPSR), as well as the literature on library-based institutional repositories. I also introduce a novel revenue model under development at the University of Maryland Libraries. I wrap up this chapter with some observations on the practical challenges and ethical problems that arise when mission-driven, public-oriented organizations seek alternative sources of revenue. My conclusion is that library-based data repositories that acquire revenue from both public and fee-for-service sources must take care to balance public and private interests in a principled, transparent way.

## From Costs to Revenue

Libraries that intend to collect, curate, and disseminate data produced by their respective research communities have to contend with a variety of costs over the life of the curated data. Costs include labor, software, hardware, network, marketing, management, and administration costs, as well as strategic costs, such as opportunity costs. Due to the prodigious rate of data growth across the research enterprise,<sup>5</sup> some repositories may find that they face persistent financial pressures related to disk usage and staff time. At the same time, the benefits of a data repository to a particular community are unevenly distributed—only some data will be reused and referenced repeatedly, and only some researchers will experience significant reputation, career, or funding benefits on account of data sharing—making it difficult to convey to university leaders and administrators that funding a data repository, and increasing that funding over time, is a worthwhile investment. In this context, it is useful for repository managers to investigate not only how to manage costs, but also new sources of income. Library-based repository managers share this concern with domain-specific repository managers, who are sufficiently alarmed about funding to have issued a “call for change” that appealed for “funding streams that are long-term, uninterrupted, and flexible.”<sup>6</sup>

Even though parent-institution funding appears to be the predominant source of revenue for library-based data repositories, we are starting to see libraries generate revenue directly from users using a few different models.\* Several libraries are already charging fees for certain repository services, and others have expressed an interest in exploring new revenue practices.<sup>7</sup> That being said, it is important to note that the examples of user fees described in this chapter are typ-

---

\* There is some precedent in academic libraries for using fee-for-service revenue models for specialized services or projects. For example, fees or charges sometimes apply to library services such as reproduction, digitization, or facilities rentals.

ically submission fees, curation assistance fees, or archiving fees (which are analogous to article processing charges in an author-pays publishing model).<sup>8</sup> None of the respondents to the survey for the 2013 ARL SPEC Kit on research data management indicated that they use access fees to generate revenue.<sup>9</sup> Curation or submission fees may be especially attractive to repository managers because many federal funding agencies in the United States permit awardees to allocate funds from their awards to support data curation and preservation.<sup>10</sup>

## Data Repository Revenue Models

Academic libraries use a variety of internal and external sources of revenue to support data repositories. In this section, I discuss six revenue models and consider some of the advantages and disadvantages of each model. I focus on revenue models for which we have public examples from library-based and, secondarily, domain-specific data repositories. The models are

1. Public or consortium
2. Freemium
3. Pay-to-play
4. Pay-if-you-can or pay-if-you-want
5. Grants
6. Outside-data

Several other widely used revenue models appear in the literature that could apply to data repositories or archives, but are probably not feasible in library-based data repositories for various reasons. (Readers should consult the works by Maron, Kitchin and colleagues, Wang and colleagues, and Erway and Rinehart for additional models.<sup>11</sup>) Among the models that appear to be infeasible are

- Selling or licensing access to data
- Selling advanced or premium data-access or -analysis features
- Advertising and corporate sponsorship
- Philanthropy

For academic librarians, these models introduce a few problems. First and foremost, librarians' mission, guiding principles, and professional ethics usually favor concepts such as equitable and open access, the protection of user privacy and confidentiality, and the attenuation of commercial interests.<sup>12</sup> Second, the commercial value of data and user traffic to library-based online data repositories varies widely, and only a limited number of repositories may be able to generate a meaningful amount of revenue from data-access or -analysis fees.<sup>13</sup> Moreover, in the United States and Europe, open-access policies increasingly require that a free copy of the data be available,<sup>14</sup> potentially undercutting any premium data products that could stimulate revenue. Third, philanthropy and other fundraising programs may generate substantial revenue, but they can also be capricious,

are subject to exhaustion, and may reflect the interests of a particular individual or group to a degree that is incompatible with the institution-wide mission of a library-based data repository.<sup>15</sup>

## *Model 1: Public or Consortium*

In a public model, an institution makes a financial investment in a library's data repository on behalf of faculty, students, and other members of the institutional community (the "public").\* The funding could come from the library, central IT, division of research, provost, or another entity. A consortium model is a variation on the public model, with funding coming from multiple members on behalf of users.

A key characteristic of public or consortium models is that end users do not usually make a direct, individual financial investment in the repository in return for service.<sup>16</sup> The public model has been described as a "free" model in the context of digital curation, but it is probably more accurate to say that the cost is subsidized in order to appear free to end users.<sup>17</sup> A consequence of this model is that the repository managers have an incentive to implement technology and provide services that maximize the benefits of the repository for the greatest number of users. When libraries use this model to support data repositories, the repositories tend to be general-purpose, domain-agnostic, publicly accessible repositories that provide roughly the same level of service to all users.

Studies of data and institutional repositories suggest that many academic libraries use the public model to cover the operating costs of repository infrastructure and related services.<sup>18</sup> In the ARL SPEC Kit on research data management services, 84 percent of respondents indicated that their data-archiving services were being funded by absorbing the cost into their respective budgets.<sup>19</sup> In addition, various surveys of institutional repositories suggest that a substantial portion of libraries use a public model to support their repositories.<sup>20</sup> Examples of library-based consortium models are rarer, but the Dataverse repository run by Scholars Portal, on behalf of a consortium of twenty-one university libraries in Ontario, Canada, is a textbook example.<sup>21</sup> The Maryland Shared Open Access Repository, a multi-institution DSpace repository operated by the University System of Maryland and Affiliated Institutions Library Consortium is similar, though not primarily marketed as a data repository.<sup>22</sup> Among domain-specific data repositories, the Inter-university Consortium for Political and Social Research (ICPSR) stands out for having about 760 members. Membership fees

---

\* I use the expression *public model* in a general sense to signify a community-focused, non-market approach to funding and not in the narrow sense of state or government funding through the redistribution of tax revenues. Thus campus libraries funded by private colleges or universities on behalf of their respective faculty and students are using a "public" model.

range from \$1,765 to \$17,400 per year (FY 2017 and 2018 rates) and depend on the member's position in the Carnegie Classification.<sup>23</sup>

The chief advantage of the public or consortium model is that it reduces the risk of market failure. In a fee-driven market, well-funded individuals or groups could use their wealth to control the design and development of a data repository, leaving other users without repository infrastructure and services. Moreover, the cost of infrastructure and services are widely distributed, so the costs for users or members may be lower. A consortium model has additional benefits because the extra revenue provided by having a large number of members could lead to improvements in technology and services that would not likely be available to the members separately.<sup>24</sup>

The public model has administrative advantages as well. For example, since there are no financial transactions with individual end users, the administrative burden related to accounting is reduced. Moreover, problems related to the distribution of revenue, such as whether the revenue belongs to the operational unit that generated it, to the library in general, or to the parent institution, are avoided.

A problem in a public or consortium model is the potential competition for funding between projects and programs that are funded through the same public or consortium arrangement. Even though a data repository may have advocates on the campus, other institutional priorities or interests can divert revenue from the repository.<sup>25</sup> Part of the problem is that the funders—in this case, usually the university and library administrators—are not typically the direct beneficiaries of the repository and may not perceive its value as clearly as users. Repository managers may find themselves engaged in ongoing advocacy and campaigning in order to sustain or increase revenue.<sup>26</sup>

For data repositories generating revenue through consortium models, a similar political problem can arise on account of the diverse interests and, perhaps, ranks of members. Members that contribute more revenue to the repository—either in direct financial or in-kind investment—may be entitled to greater influence over the priorities of the repository. Thus the effort to manage the interests of members may become a significant administrative cost for the repository.<sup>27</sup>

## *Model 2: Freemium*

The freemium revenue model divides the product or service into distinct service levels: (1) a free, basic level that provides limited service and (2) one or more fee-for-service, premium levels that provide additional services, performance, or support.<sup>28</sup> The freemium model is fairly common in cloud-based consumer applications, and Dropbox and Spotify are characteristic examples. It is interesting to note that the levels of service may encourage different social or cultural practices. For example, GitHub provides free software repositories to publicly available



projects and charges a fee for repositories used by private projects, effectively using the revenue model to subsidize (and therefore encourage) open-source culture. The freemium model may also be used to support free services for non-commercial users and extract revenue from commercial users or licensees.<sup>29</sup> The main challenge for an enterprise operating on a freemium model is to stimulate enough users to purchase the premium level to support, and ideally exceed, the costs associated with providing the free level.<sup>30</sup>

A small number of data repositories use a freemium model to generate revenue. Table 8.1 provides evidence from three library-based data repositories and one domain-specific repository.

These examples illustrate that data repositories are generating revenue by charging a fee for extra data storage (disk usage by gigabyte) or extra curation assistance. Some incorporate a retention period restriction (e.g., the Purdue University Research Repository) to further manage costs. In the ARL SPEC Kit on research data management, 14 percent of respondents indicate that they charge a fee to data producers and the de-identified narrative responses suggest that several of them use a freemium model based on disk usage.<sup>31</sup> It is worth noting, though, that these examples may not be perfectly analogous to commercial enterprises that operate on a freemium model because data repositories likely have other revenue sources—probably public or consortium models—to support the “free” level of infrastructure and services. In other words, it is unlikely that library-based data repositories would go out of business if they failed to convert users from the free to premium level. These libraries and archives appear to use the freemium model to offset some or all of the costs associated with resource-intensive services.

The freemium model has a number of advantages for repository managers. When the free level is subsidized by, for example, a public model, the repository can support all data producers to some extent, which protects the data repository service against market failure. At the same time, fees for extra data storage can protect the repository against rising infrastructure and management costs related to increasing disk usage. Fees for extra curation assistance may recover some or all of the personnel costs associated with verifying and improving the quality, understandability, and accessibility of data. These fees may also provide a mechanism for capturing revenue from the portion of grant awards that may be allocated to data management.

Data producers may benefit from the freemium model by having a basic service that can guarantee some degree of data curation and preservation as well as the option to purchase additional services at predictable cost. On the other hand, when looking at the freemium model from a data producer’s perspective, there are clear disadvantages. In the case of extra curation services, data producers may perceive that improvements in data quality, understandability, and accessibility are not worth the expense, particularly if they are submitting data to the repository only to satisfy the regulations of their respective funding entities. If

the free level of service is “good enough” for most cases, then it is unlikely that the repository will persuade users to purchase the premium level.<sup>32</sup> In addition, in the case of fees for extra data storage, the freemium model may penalize data producers who collect or generate large volumes of data but are not well-funded.

**TABLE 8.1**  
Freemium Revenue Practices at Data Repositories

Repository	Free Level	Fee-for-Service Level
<b>University of Nebraska-Lincoln Data Repository<sup>a</sup></b> University of Nebraska-Lincoln Libraries	Submissions up to 50GB.	One-time fee for submissions over 50GB in tiers based on disk usage.
<b>Purdue University Research Repository<sup>b</sup></b> Purdue University	Data publications up to 1GB or, for grant awardees, up to 10GB.	“Extra publication space” priced at \$14.30/GB for 10 years of data publication.
<b>JHU Data Archive<sup>c</sup></b> Johns Hopkins University	“Small Data Collections Archiving Service” for submissions up to 20GB for 5 years. Additional fees apply to large volumes of data or longer retention period.	“Large Data Collections Single Grant Archiving Service” is priced at 2% of total direct costs on grant for up to 2TB for 5 years. Includes a variety of curation assistance.
<b>openICPSR<sup>d</sup></b> Inter-university Consortium for Political and Social Research	“Self-Deposit Package” for ICPSR members (\$600 for nonmembers).	“Professional Curation Package” (price not disclosed). Service includes “full metadata generation and a bibliography search, stat package conversion, and user support.” <sup>e</sup>

Note: Published pricing information valid on February 19, 2016.

a. University of Nebraska-Lincoln Libraries, “Submitting Data,” University of Nebraska-Lincoln Data Repository, accessed November 22, 2015, <https://dataregistry.unl.edu/researchers.html>.

b. Purdue University Research Repository, “PURR Project Space Allocation and Pricing,” accessed November 22, 2015, <https://purrr.purdue.edu/aboutus/pricing>.

c. Johns Hopkins University Data Management Services, “Archiving Services We Offer,” Johns Hopkins University Data Management Services, accessed July 26, 2016, <http://dmp.data.jhu.edu/preserve-share-research-data/archiving-services-we-offer/>.

d. Inter-university Consortium for Political and Social Research, “Plans & Pricing,” openICPSR, accessed November 22, 2015, <https://www.openicpsr.org/openicpsr/pricing>.

e. Inter-university Consortium for Political and Social Research, “FAQs,” openICPSR, accessed November 22, 2015, <https://www.openicpsr.org/openicpsr/faqs>.

## Model 3: Pay-to-Play

In the pay-to-play model, data repositories charge a fee for every submission. This is not a popular model for academic libraries, but as recently as 2015 Princeton University's DataSpace and Johns Hopkins University's JHU Data Archive charged a minimum fee for submissions. Princeton University charged a one-time fee of \$0.006 per MB and \$0.60 minimum charge per submission. Johns Hopkins University charged \$1,600 for submissions in a "Small Data Collections Archiving Service" covering up to 20GB for five years.<sup>33</sup> Both appeared to have waived those fees by early 2016, leaving non-library-based repositories as the best sources of evidence about this revenue model. Table 8.2 provides examples from the Dryad and tDAR (the Digital Archaeological Record) data repositories. Open Context, an archaeology data repository, also charges a submission fee, but discloses its fees only on an estimate basis.<sup>34</sup>

**TABLE 8.2**  
Pay-to-Play Revenue Practices at Data Repositories

Repository	Fee
<b>Dryad Digital Repository<sup>a</sup></b> Dryad	"Data Publishing Charge" for individuals is priced at \$120USD per data package. Additional charges apply to data packages that exceed 20GB. Volume and country-of-origin discounts are available. Some "basic checks" are included for quality control.
<b>tDAR<sup>b</sup></b> Digital Antiquity	\$10/file for 1–99 files and \$5/file for more than 100 files. Each file is entitled to 10MB. A basic curation service including metadata creation and quality control is \$90/hour. An enhanced service including programming and project management is \$180/hour.

Note: Published pricing information for Dryad valid on February 19, 2016. Information for tDAR valid on March 25, 2016.

a. Dryad Digital Repository, "Payment Plans and Data Publishing Charges," last revised January 5, 2016, <http://datadryad.org/pages/payment>.

b. Digital Archaeological Record, "Pricing," accessed March 26, 2016, <https://core.tdar.org/cart/add>.

In these examples, as with the freemium model, fees are applied on the basis of disk usage and curation assistance. Dryad has expressed its intention to become a "financially independent non-profit organization,"<sup>35</sup> so its pricing scheme will become an indicator about the pricing required to curate and disseminate data, code, and other research products on a self-sustaining basis.

For repository managers, the pay-to-play model provides a mechanism for guaranteeing some degree of cost recovery, discouraging excessive consumption

of services, and capturing funding from grants. It also provides an interesting signal about the apparent financial value of the data repository. When faculty and students deposit data into publicly funded repository services that appear to be free, it can be hard to tell how much of that demand should be attributed to the fact that the repository is free and how much to the features and benefits of the repository. When users pay directly for repository services, the repository managers have evidence that the repository provides features and benefits equal to or greater than the price.

The chief disadvantage of the pay-to-play model in a library-based data repository is that the price may drive away data producers, even those who are well-funded. Moreover, grant-funded researchers may wonder why the cost of repository services is not covered by indirect or facilities-and-administration costs extracted from their awards.<sup>36</sup> Another problem is that some general-purpose data repositories, notably Harvard Dataverse, figshare, Zenodo, and Open Science Framework, are currently free to data producers from any institution for submissions that fall within certain limits, creating competition for users. For these reasons, a key task for repository managers considering a pay-to-play model is to offer services that have a commensurate or, ideally, higher value than the price.

## *Model 4: Pay-if-You-Can or Pay-if-You-Want*

The pay-if-you-can model provides free services to all users but introduces fees for those users who have the means to pay. It is loosely related to the pay-what-you-want model that has been used to sell various consumer products, in which some users may elect to pay zero while other users may elect to pay more than zero for the same product or service.<sup>36</sup> The two models are related because it is possible that all users will pay zero. For that reason, this model should have another source of revenue to guarantee ongoing operations.

I have not encountered a library-based data repository that uses the pay-if-you-can model, but PANGAEA, a data repository for earth and environmental science based in Europe, requests a contribution of, on average, €300 from depositors whose funding includes money for publication costs.<sup>37</sup> It is not clear how a data repository would enforce this condition without devoting some effort to verification, so this version of the pay-if-you-can model may depend on the good faith of users.

---

\* In the United States, universities and colleges are usually entitled to extract funds from grant awards to cover indirect or facilities-and-administration costs incurred by awardees. These are typically general campus expenses such as buildings, equipment, technology, grant administration, libraries, and so on.

In comparison to the freemium model, the pay-if-you-can and pay-what-you-want models reduce the risk of penalizing users who for one reason or other—perhaps disk usage—trigger premium-level services but are nevertheless not well-funded. However, the challenge for data repository managers implementing a pay-if-you-can model is designing conditions that are understood by all users to be fair. For example, PANGAEA's condition may not work well for a library-based data repository that has some institutional funding because grant awardees may feel that they are being unfairly penalized. One way to reduce the risk of unfair or asymmetric conditions is to remove all conditions and use a pay-what-you-want model that relies on voluntary donations. Open Context has a mechanism for receiving donations in addition to other revenue sources.<sup>38</sup> Pay-what-you-want models have had some successes in various sectors of the economy, and some incentive mechanisms have been proven—such as promising to direct a portion of revenue to charity—but the evidence is largely from popular consumer goods.<sup>39</sup>

## *Model 5: Grants*

It is evident that a small number of library-based data repositories have received grants to support repository development or operation. In the ARL SPEC Kit on research data management services, 24 percent of respondents indicated that some form of grant funding pays or paid for data archiving, though it is not clear from that survey whether those are grants for infrastructure development or data curation allocations from research grant budgets.<sup>40</sup> Studies of institutional repositories suggest that grant funding has contributed to repository revenues in a small number of cases.<sup>41</sup> Evidence from surveys suggests that grants have typically been used to cover startup costs more often than ongoing operating costs.<sup>42</sup>

Grants are useful as sources of revenue because they enable repository managers to accelerate development. However, they come with a critical drawback: they are typically short-duration revenue and not intended to cover long-term operating or maintenance costs.<sup>43</sup> Increasingly, grant proposals for digital curation projects have to articulate how operating costs will be covered once the award ends, so in practice grants are best employed as part of a hybrid revenue model.<sup>44</sup>

## *Model 6: Outside-Data*

The outside-data model is a novel revenue practice being explored by the University of Maryland Libraries. It is similar to a documented business model known as “make more of it.”<sup>45</sup> In that model, “know-how or other resources are offered to outside companies in addition to being used in-house. In this manner, ‘slack’ resources help to add revenue on top of the core value proposition’s returns.”<sup>46</sup>

Amazon Web Services and Google Cloud Platform are examples of this model.<sup>47</sup> Amazon and Google lease elements of the technology that they use to operate their respective businesses—their storage, computing resources, APIs, and so on—to outside individuals and organizations to use for independent purposes. Maron (2014) described an analogous model, called the “consulting” model, whereby a digital curation organization may sell its expertise to clients who are not its customary clients.<sup>48</sup>

At the University of Maryland Libraries, the outside-data model is being examined as a mechanism for generating revenue by using the Libraries’ repository capabilities to curate, preserve, and, where applicable, provide access to non-research data, such as operational or administrative data. This data may be produced by campus units, private sector organizations, government agencies, or other entities. While this data may be subject to access policies and other compliance requirements, it is not necessarily subject to the open-access policies that affect research data. There are a few assumptions behind this model:

1. Curation and preservation of non-research data are not in the library’s mission and not supported by the library’s traditional revenue (a public model).
2. The data owner (the client) would have paid a vendor or supplier for the same or equivalent services. Iron Mountain is an example of a commercial firm providing “information management” services that are an amalgam of records management and digital curation.
3. When accepted, the non-research and research data will share much of the same technology infrastructure, so the marginal cost of curating and preserving the non-research data is low (but not zero).
4. The client acknowledges that the library may reuse technology and services funded by the non-research data to benefit research data collections.

For example, a university counsel could use the library’s data preservation and access capabilities to ensure the authenticity, integrity, and long-term understandability of the institution’s legal records. For these records, access policies and procedures would likely be more restrictive than for research data, but much of the library’s infrastructure and expertise would still be relevant and useful.

A distinctive benefit for the library from curating this kind of data—aside from revenue—is that any improvements to infrastructure, processes, procedures, or knowledge that are funded or stimulated by the non-research data curation may be applied to research data curation. At the same time, the client organization benefits from access to the digital preservation and records management expertise of the library. The cost to the client may also be lower since the library may not seek a margin as large as that sought by a private-sector vendor or supplier. The library may simply seek enough margin to support ongoing investment in the data repository.

An objection to this approach is that most academic libraries do not have surplus resources, especially staff time, to spare on non-research or non-academic data curation and preservation. In addition, running such an operation may

distract the repository from its main mission and introduce ethical conflicts.<sup>49</sup> The general approach to solving this problem is to reduce the marginal cost of curating non-research data—including the costs of administration and management—to as close to zero as possible. At one extreme, this may involve building a separate, self-sustaining organization with its own personnel and administration to manage the non-research data, which is a substantially complex operation.

## Common Challenges Associated with Revenue Practices

Libraries that wish to implement new revenue models, particularly direct revenue models involving fee-for-service practices, will encounter a variety of hurdles. Based on experiences at the University of Maryland Libraries and building on insights from the literature, several challenges stand out as especially problematic. These have to do with pricing, motivating demand, operational efficiencies, and managing potential conflicts of interest.

Charging fees for particular services introduces the problem of pricing: what exactly should a data repository charge for disk usage, curation assistance, or another product? Answering this question requires a detailed understanding of short- and long-term repository costs.\* Once costs are understood, the repository managers have to decide whether to seek profits, recover costs, or absorb overruns. Profit seeking is generally anathema in mission-driven, public-oriented organizations, and respondents to the 2015 ARL SPEC Kit survey on research data management mentioned only cost recovery or chargeback.<sup>50</sup> In addition, profit seeking may be prohibited between entities in the same institution, such as libraries and research teams. Nevertheless, repository managers have to consider how growth will be funded. Some additional revenue is probably necessary to develop new technologies and services, even in a pure cost-recovery context.

Repositories that depend completely or substantially on direct revenue from data producers also face the problem of motivating demand for curation services. Even if the service model and pricing are attractive to data producers, a steady supply of clients is probably necessary to cover annual operating costs. The marketing effort required to generate demand should not be underestimated and, in all likelihood, success will require that the library devote more resources to marketing, potentially neutralizing the benefits of the revenue generated.

Repositories may also need to develop competence in business development and client management in order to attract and retain clients. This set of practic-

---

\* See literature references in Notes 1, 2, & 3.



es is not typically part of the training or professional development in academic librarianship and may have to be acquired (possibly at some cost). At a more mundane level, revenue models that involve direct financial transactions between users and repository managers require accounting processes for estimating, invoicing, accounts receivable, credit notes, payments, receipts, and so on. While these processes usually exist in academic libraries, they may not have the efficiency of a retail operation. Even accepting a credit card as payment can be improbable for some libraries, though turnkey retail payment systems such as Square can make this easier. Introducing these processes into a data repository may also create an administrative burden for managers.<sup>51</sup>

Finally, the greatest problem for library-based data repositories that receive revenues through a hybrid approach that involves both a public model and a fee-for-service or other non-public model is to manage conflicts between the public and private interests. If a university or college understands that it funds 100 percent of a data repository's expenditures—which is probably the situation at many academic libraries—then users may think, rightfully so, that the repository services are fully and already funded. In this situation, fee-for-service activities would appear to buy services that have already been purchased. To act responsibly, data repository managers who want to generate revenue from combinations of public and fee-for-service models have to articulate precisely to what extent the infrastructure and services, especially staff time, are covered by each source of revenue.

## Conclusion

The challenges associated with finding and implementing new sources of revenue should not deter data repository managers and library administrators from investigating such opportunities. Evidence from various studies suggest that quite a few academic libraries recognize the need for additional revenue sources to ensure the financial sustainability of their respective data repositories. Several are likely experimenting with hybrid models that combine public funding and fee-for-service models.<sup>52</sup> Yet, as this chapter has shown, new revenue models come with ethical tensions and practical difficulties. Revenue models that have the best chance of success are those that (1) reflect the apparent value of the repository and (2) have the lowest risk of creating conflicts between public and private interests. The former is a question of service design, and the latter a question of business modeling. One way to resolve the conflict between public and private interests is to operate on either a public or fee-for-service model exclusively. Another way is to use a freemium model, but with an obvious separation of capabilities between the free and premium levels of services. The outside-data model also presents an interesting, though complex, opportunity to generate revenue by offering existing

data repository infrastructure and services to users whose interests are orthogonal to the research enterprise but whose data preservation needs are similar.

## Notes

1. For cost models see 4C Collaboration to Clarify the Costs of Curation, “Summary of Cost Models,” accessed November 22, 2015, <http://4cproject.eu/summary-of-cost-models>.
2. For an introduction to the literature on costs see Butch Lazorchak, “A Digital Asset Sustainability and Preservation Cost Bibliography,” *The Signal: Digital Preservation* (blog), June 26, 2012, <http://blogs.loc.gov/digitalpreservation/2012/06/a-digital-asset-sustainability-and-preservation-cost-bibliography/>.
3. For cost models see 4C Collaboration to Clarify the Costs of Curation, “Summary of Cost Models”; for an introduction to the literature on costs see Lazorchak, “A Digital Asset Sustainability and Preservation Cost Bibliography”; an effort is underway to share curation costs at Curation Costs Exchange, accessed November 22, 2015, <http://www.curationexchange.org/>.
4. Nancy Maron, *A Guide to the Best Revenue Models and Funding Sources for Your Digital Resources* (New York: Ithaca S+R, 2014), <http://www.sr.ithaka.org/publications/a-guide-to-the-best-revenue-models-and-funding-sources-for-your-digital-resources/>; Rob Kitchin, Sandra Collins, and Dermot Frost, “Funding Models for Open Access Digital Data Repositories,” *Online Information Review* 39, no. 5 (September 14, 2015): 664–81, doi:10.1108/OIR-01-2015-0031; David Wang, Stephan Strodl, Tomasz Miksa, Ulla Bøgvad Kejser, Miguel Ferreira, Jose Borbinha, Diogo Proença, et al., “From Costs to Business Models,” Collaboration of Clarify the Cost of Curation, 4C Project, 2015, <http://www.4cproject.eu/d4-5-from-costs-to-business-models>; Ricky Erway and Amanda Rinehart, *If You Build It, Will They Fund? Making Research Data Management Sustainable* (Dublin, OH: OCLC Research, 2016), <http://www.oclc.org/content/dam/research/publications/2016/oclcresearch-making-research-data-management-sustainable-2016.pdf>.
5. Gordon Bell, Tony Hey, and Alex Szalay, “Beyond the Data Deluge,” *Science* 323, no. 5919 (March 6, 2009): 1297–98, doi:10.1126/science.1170411; Gary King, “Ensuring the Data-Rich Future of the Social Sciences,” *Science* 331, no. 6018 (February 11, 2011): 719–21, doi:10.1126/science.1197872.
6. Inter-university Consortium for Political and Social Research, “Sustaining Domain Repositories for Digital Data: A Call for Change from an Interdisciplinary Working Group of Domain Repositories,” June 24–25, 2013, <http://www.icpsr.umich.edu/files/ICPSR/pdf/DomainRepositoriesCTA16Sep2013.pdf>.
7. David Fearon Jr., Betsy Gunia, Barbara E. Pralle, Sherry Lake, and Andrew L. Sallans, *Research Data Management Services: SPEC Kit 334* (Washington, DC: Association of Research Libraries, 2013), 50–51, <http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/>.
8. Maron, *A Guide to the Best Revenue Models*, 24–30.
9. Fearon et al., *Research Data Management Services*, 50.
10. Erway and Rinehart, *If You Build It, Will They Fund?* 4–5.
11. Maron, *A Guide to the Best Revenue Models*; Kitchin, Collins, and Frost, “Funding Models for Open Access Digital Data Repositories”; Wang et al., “From Costs to Business

- Models”; Erway and Rinehart, *If You Build It, Will They Fund?*
12. Maron, *A Guide to the Best Revenue Models*, 11–12, 68; Kitchin, Collins, and Frost, “Funding Models for Open Access Digital Data Repositories,” 669–672; Wang et al., “From Costs to Business Models,” 19–20; American Library Association, *Code of Ethics of the American Library Association* (Chicago: American Library Association, 1939; amended <http://www.ala.org/advocacy/proethics/codeofethics/codeethics>).
  13. Kitchin, Collins, and Frost, “Funding Models for Open Access Digital Data Repositories,” 671; Maron, *A Guide to the Best Revenue Models*, 17.
  14. John P. Holdren, “Increasing Access to the Results of Federally Funded Scientific Research,” Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013, [https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf); Wilma Mossink, Magchiel Bijsterbosch, and Joeri Nortier, *European Landscape Study of Research Data Management* (Utrecht, NL: SURFfoundation, 2013), [http://www.clarin.nl/sites/default/files/SIM4RDM%20landscape%20report\\_1.pdf](http://www.clarin.nl/sites/default/files/SIM4RDM%20landscape%20report_1.pdf); Research Councils UK, “RCUK Common Principles on Data Policy,” 2015, <http://www.rcuk.ac.uk/research/datapolicy/>.
  15. Kitchin, Collins, and Frost, “Funding Models for Open Access Digital Data Repositories,” 671.
  16. Wang et al., “From Costs to Business Models,” 17.
  17. *Ibid.*, 16.
  18. Fearon et al., *Research Data Management Services*; Erway and Rinehart, *If You Build It, Will They Fund?*; Soo Young Rieh, Karen Markey, Beth St. Jean, Elizabeth Yakel, and Jihyun Kim, “Census of Institutional Repositories in the U.S.: A Comparison across Institutions at Different Stages of IR Development,” *D-Lib Magazine* 13, no. 11/12 (November 2007), doi:10.1045/november2007-rieh; Karen Bjork, David Isaak, and Kay Vyhnanek, “The Changing Roles of Repositories: Where We Are and Where We Are Headed” (presentation, Oregon Library Association/Washington Library Association Conference, Vancouver, WA, April 25, 2013), Library Faculty Publications and Presentations, [http://pdxscholar.library.pdx.edu/ulib\\_fac/84](http://pdxscholar.library.pdx.edu/ulib_fac/84); Charles W. Bailey Jr., Karen Coombs, Jill Emery, Anne Mitchell, Chris Morris, Spenser Simons, and Robert Wright, *Institutional Repositories, SPEC Kit 292* (Washington, DC: Association of Research Libraries, 2006), <http://publications.arl.org/Institutional-Repositories-SPEC-Kit-292/>.
  19. Fearon et al., *Research Data Management Services*.
  20. Rieh et al., “Census of Institutional Repositories in the U.S.”; Bjork, Isaak, and Vyhnanek, “The Changing Roles of Repositories”; Bailey et al., *Institutional Repositories*, 16.
  21. Scholars Portal, “Dataverse at Scholars Portal: Home,” 2016, <http://guides.scholarsportal.info/dataverse>.
  22. University System of Maryland and Affiliated Institutions Library Consortium, Maryland Shared Open Access Repository homepage, accessed March 25, 2016, <https://mdsoar.org/>.
  23. Inter-university Consortium for Political and Social Research, “Membership in ICPSR,” accessed March 26, 2016, <https://www.icpsr.umich.edu/icpsrweb/content/membership/index.html>; Inter-university Consortium for Political and Social Research, “How to Become a Member or Subscribe to ICPSR,” accessed March 26, 2016, <https://www.icpsr.umich.edu/icpsrweb/content/membership/join.html>.

24. See, for example, California Digital Library, *Budget Background 2014–2015* (University of California Libraries, November 18, 2013), 2, [http://libraries.universityofcalifornia.edu/groups/files/slasiac/docs/CDL\\_Budget\\_Background\\_2014-2015.pdf](http://libraries.universityofcalifornia.edu/groups/files/slasiac/docs/CDL_Budget_Background_2014-2015.pdf).
25. Maron, *A Guide to the Best Revenue Models*, 48, 51.
26. *Ibid.*, 51.
27. This kind of management overhead appears in a number of models in Maron, *A Guide to the Best Revenue Models*.
28. Oliver Gassmann, Karolin Frankenberger, and Michaela Csik, *The Business Model Navigator* (Harlow, UK: Pearson Education, 2014), 166–168.
29. Maron, *A Guide to the Best Revenue Models*, 43–44; Kitchin, Collins, and Frost, “Funding Models for Open Access Digital Data Repositories,” 669.
30. Gassmann, Frankenberger, and Csik, *The Business Model Navigator*, 166.
31. In the ARL SPEC Kit on research data management (Fearon et al., *Research Data Management Services*), 14 percent of respondents indicate that they charge a fee to data producers, and the de-identified narrative responses suggest that several of them use a freemium model based on disk usage.
32. Maron, *A Guide to the Best Revenue Models*, 41–42.
33. Princeton University, “DataSpace at Princeton University,” August 18, 2015, <https://web.archive.org/web/20150818071415/http://dataspace.princeton.edu/jspui/about/home.jsp>; Johns Hopkins University Data Management Services, “Archiving Services We Offer,” September 5, 2015, <https://web.archive.org/web/20150905163236/http://dmp.data.jhu.edu/preserve-share-research-data/archiving-services-we-offer/>.
34. Open Context, “Publishing and Archiving Costs,” accessed February 19, 2016, <http://opencontext.org/about/estimate>.
35. tjvision, “NSF Provides Further Support to Dryad,” *Dryad News and Views* (blog), Dryad Digital Repository, March 1, 2012, <http://blog.datadryad.org/2012/03/01/nsf-provides-further-support-to-dryad/>.
36. Gassmann, Frankenberger, and Csik, *The Business Model Navigator*, 249–251.
37. PANGAEA, “Submit Data to PANGAEA,” accessed November 20, 2015, <http://www.pangaea.de/submit/>.
38. Alexandria Archive Institute, “Support Our Work,” accessed February 19, 2016, <http://alexandriaarchive.org/contribute/?ref=opencontext>.
39. Hyunkyung Jang and Wujin Chu, “Are Consumers Acting Fairly Toward Companies? An Examination of Pay-What-You-Want Pricing,” *Journal of Macromarketing* 32, no. 4 (July 16, 2012): 349, doi:10.1177/0276146712448193; Ayelet Gneezy, Uri Gneezy, Gerhard Riener, and Leif D. Nelson, “Pay-What-You-Want, Identity, and Self-Signaling in Markets,” *Proceedings of the National Academy of Sciences* 109, no. 19 (May 8, 2012): 7236–40, doi:10.1073/pnas.1120893109.
40. Fearon et al., *Research Data Management Services*, 50.
41. Rieh et al., “Census of Institutional Repositories in the U.S.”; Bjork, Isaak, and Vyhnanek, “The Changing Roles of Repositories.”
42. Bailey et al., *Institutional Repositories*, 57; Bjork, Isaak, and Vyhnanek, “The Changing Roles of Repositories.”
43. Kitchin, Collins, and Frost, “Funding Models for Open Access Digital Data Repositories,” 672.
44. Maron, *A Guide to the Best Revenue Models*, 71.
45. Gassmann, Frankenberger, and Csik, *The Business Model Navigator*, 216–220.

46. Ibid., 217.
47. Ibid., 218.
48. Maron, *A Guide to the Best Revenue Models*, 31–35.
49. Ibid., 34.
50. Fearon et al., *Research Data Management Services*, 51; Maron, *A Guide to the Best Revenue Models*, 11–12.
51. Maron, *A Guide to the Best Revenue Models*.
52. Kitchin, Collins, and Frost, “Funding Models for Open Access Digital Data Repositories,” 678.

## Bibliography

- Alexandria Archive Institute. “Support Our Work.” Accessed February 19, 2016. <http://alexandriaarchive.org/contribute/?ref=opencontext>.
- American Library Association. *Code of Ethics of the American Library Association*. Chicago: American Library Association, 1939; amended 1981, 1995, 2008. <http://www.ala.org/advocacy/proethics/codeofethics/codeethics>.
- Bailey, Charles W. Jr., Karen Coombs, Jill Emery, Anne Mitchell, Chris Morris, Spenser Simons, and Robert Wright. *Institutional Repositories, SPEC Kit 292*. Washington, DC: Association of Research Libraries, 2006. <http://publications.arl.org/Institutional-Repositories-SPEC-Kit-292/>.
- Bell, Gordon, Tony Hey, and Alex Szalay. “Beyond the Data Deluge.” *Science* 323, no. 5919 (March 6, 2009): 1297–98. doi:10.1126/science.1170411.
- Bjork, Karen, David Isaak, and Kay Vyhnanek. “The Changing Roles of Repositories: Where We Are and Where We Are Headed.” Presentation, Oregon Library Association/Washington Library Association Conference, Vancouver, WA, April 25, 2013. Library Faculty Publications and Presentations. [http://pdxscholar.library.pdx.edu/ulib\\_fac/84](http://pdxscholar.library.pdx.edu/ulib_fac/84).
- California Digital Library. *Budget Background 2014–2015*. University of California Libraries, November 18, 2013. [http://libraries.universityofcalifornia.edu/groups/files/slasiac/docs/CDL\\_Budget\\_Background\\_2014-2015.pdf](http://libraries.universityofcalifornia.edu/groups/files/slasiac/docs/CDL_Budget_Background_2014-2015.pdf).
- Curation Costs Exchange homepage. Accessed November 22, 2015. <http://www.curationexchange.org/>.
- Digital Archaeological Record. “Pricing.” Accessed March 26, 2016. <https://core.tdar.org/cart/add>.
- Dryad Digital Repository. “Payment Plans and Data Publishing Charges.” Last revised January 5, 2016. <http://datadryad.org/pages/payment>.
- Erway, Ricky, and Amanda Rinehart. *If You Build It, Will They Fund? Making Research Data Management Sustainable*. Dublin, OH: OCLC Research, 2016. <http://www.oclc.org/content/dam/research/publications/2016/oclcresearch-making-research-data-management-sustainable-2016.pdf>.
- Fearon, David Jr., Betsy Gunia, Barbara E. Pralle, Sherry Lake, and Andrew L. Sallans. *Research Data Management Services: SPEC Kit 334*. Washington, DC: Association of Research Libraries, 2013. <http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/>.
- 4C Collaboration to Clarify the Costs of Curation. “Summary of Cost Models.” Accessed November 22, 2015. <http://4cproject.eu/summary-of-cost-models>.

- Gassmann, Oliver, Karolin Frankenberger, and Michaela Csik. *The Business Model Navigator: 55 Models That Will Revolutionise Your Business*. Harlow, UK: Pearson Education, 2014.
- Gneezy, Ayelet, Uri Gneezy, Gerhard Riener, and Leif D. Nelson. "Pay-What-You-Want, Identity, and Self-Signaling in Markets." *Proceedings of the National Academy of Sciences* 109, no. 19 (May 8, 2012): 7236–40. doi:10.1073/pnas.1120893109.
- Holdren, John P. "Increasing Access to the Results of Federally Funded Scientific Research." Memorandum for the Heads of Executive Departments and Agencies, Office of Science and Technology Policy, Executive Office of the President, February 22, 2013. [https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf).
- Inter-university Consortium for Political and Social Research. "FAQs." openICPSR. Accessed December 20, 2016. <https://www.openicpsr.org/openicpsr/faqs>.
- . "How to Become a Member or Subscribe to ICPSR." Accessed March 26, 2016. <https://www.icpsr.umich.edu/icpsrweb/content/membership/join.html>.
- . "Membership in ICPSR." Accessed March 26, 2016. <https://www.icpsr.umich.edu/icpsrweb/content/membership/index.html>.
- . "Plans & Pricing." openICPSR. Accessed December 20, 2016. <https://www.openicpsr.org/openicpsr/pricing>.
- . "Sustaining Domain Repositories for Digital Data: A Call for Change from an Interdisciplinary Working Group of Domain Repositories." June 24–25, 2013. <http://www.icpsr.umich.edu/files/ICPSR/pdf/DomainRepositoriesCTA16Sep2013.pdf>.
- Jang, Hyunkyung, and Wujin Chu. "Are Consumers Acting Fairly toward Companies? An Examination of Pay-What-You-Want Pricing." *Journal of Macromarketing* 32, no. 4 (July 16, 2012): 348–60. doi:10.1177/0276146712448193.
- Johns Hopkins University Data Management Services. "Archiving Services We Offer." September 5, 2015. <https://web.archive.org/web/20150905163236/http://dmp.data.jhu.edu/preserve-share-research-data/archiving-services-we-offer/>.
- King, Gary. "Ensuring the Data-Rich Future of the Social Sciences." *Science* 331, no. 6018 (February 11, 2011): 719–21. doi:10.1126/science.1197872.
- Kitchin, Rob, Sandra Collins, and Dermot Frost. "Funding Models for Open Access Digital Data Repositories." *Online Information Review* 39, no. 5 (September 14, 2015): 664–81. doi:10.1108/OIR-01-2015-0031.
- Lazorchak, Butch. "A Digital Asset Sustainability and Preservation Cost Bibliography," *The Signal: Digital Preservation* (blog), June 26, 2012. <http://blogs.loc.gov/digitalpreservation/2012/06/a-digital-asset-sustainability-and-preservation-cost-bibliography/>.
- Maron, Nancy. *A Guide to the Best Revenue Models and Funding Sources for Your Digital Resources*. New York: Ithaka S+R, 2014. <http://www.sr.ithaka.org/publications/a-guide-to-the-best-revenue-models-and-funding-sources-for-your-digital-resources/>.
- Mossink, Wilma, Magchiel Bijsterbosch, and Joeri Nortier. *European Landscape Study of Research Data Management*. Utrecht, NL: SURFfoundation, 2013. [http://www.clarin.nl/sites/default/files/SIM4RDM%20landscape%20report\\_1.pdf](http://www.clarin.nl/sites/default/files/SIM4RDM%20landscape%20report_1.pdf).
- Open Context. "Publishing and Archiving Costs." Accessed February 19, 2016. <http://opencontext.org/about/estimate>.
- PANGAEA. "Submit Data to PANGAEA." Accessed November 20, 2015. <http://www.pangaea.de/submit/>.

- Princeton University. "DataSpace at Princeton University." August 18, 2015. <https://web.archive.org/web/20150818071415/http://dataspace.princeton.edu/jspui/about/home.jsp>.
- Purdue University Research Repository. "PURR Project Space Allocation and Pricing." Accessed November 22, 2015. <https://purr.purdue.edu/aboutus/pricing>.
- Research Councils UK. "RCUK Common Principles on Data Policy." 2015. <http://www.rcuk.ac.uk/research/datapolicy/>.
- Rieh, Soo Young, Karen Markey, Beth St. Jean, Elizabeth Yakel, and Jihyun Kim. "Census of Institutional Repositories in the U.S.: A Comparison across Institutions at Different Stages of IR Development." *D-Lib Magazine* 13, no. 11/12 (November 2007). doi:10.1045/november2007-rieh.
- Scholars Portal. "Dataverse at Scholars Portal: Home." 2016. <http://guides.scholarsportal.info/dataverse>.
- tjvision. "NSF Provides Further Support to Dryad." *Dryad News and Views* (blog). Dryad Digital Repository. March 1, 2012. <http://blog.datadryad.org/2012/03/01/nsf-provides-further-support-to-dryad/>.
- University of Nebraska-Lincoln Libraries. "Submitting Data." University of Nebraska-Lincoln Data Repository. Accessed November 22, 2015. <https://dataregistry.unl.edu/researchers.html>.
- University System of Maryland and Affiliated Institutions Library Consortium. Maryland Shared Open Access Repository homepage. Accessed March 25, 2016. <https://md-soar.org/>.
- Wang, David, Stephan Strodl, Tomasz Miksa, Ulla Bøgvad Kejser, Miguel Ferreira, Jose Borbinha, Diogo Proença, et al. "From Costs to Business Models." Collaboration of Clarify the Cost of Curation. 4C Project, 2015. <http://www.4cproject.eu/d4-5-from-costs-to-business-models>.







## CHAPTER 9\*

# Current Outreach and Marketing Practices for Research Data Repositories

*Katherine J. Gerwig*

Libraries are increasingly creating data services to assist researchers with meeting funder data-sharing requirements and the stewardship of their research data. As librarians take the lead in establishing research data repositories and other data services, they are also tasked with soliciting researchers to deposit their data, into either the local data repository or an appropriate disciplinary repository. Research data repositories present an opportunity for librarians to leverage their expertise in curation, outreach, and preservation while strengthening their long-standing relationships with academic departments in order to implement robust repositories that satisfy the needs of their communities. Data repositories require promotion and outreach activities that not only create awareness of the repository but also inform researchers of the benefits of data sharing while addressing the needs and concerns of multiple research cultures. An examination of the promotional activities of data repositories from the perspective of those closely associated with the repository or data services can provide information to develop more effective repository outreach practices.

---

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

Institutional repositories (IRs) got off to a rough start. One reason for the low submission rates to IRs is their basis in providing a solution to the rising costs of journals directly impacting library budgets and not a problem directly affecting researchers.<sup>1</sup> In contrast, Choudhury's 2008 *Library Trends* article acknowledged data repositories as largely driven by the needs of researchers. Furthermore, he posited that we should view the IR and data repositories as parts of a wider infrastructure emerging to support researchers.<sup>2</sup> Implications of viewing the IR and data repository in this way raise the question: How are institutions promoting the IR and data repositories, together or separately?

Similar to promoting IRs, promoting data repositories involves forming partnerships with campus departments and reaching out to researchers and students. In fact, many institutions use the IR to house data.<sup>3</sup> IRs house published articles, theses and dissertations, and gray literature that can be harvested from the web or submitted by graduate students eager to make their work available. Since data is not typically published and remains in the sole custody of the researcher, direct action on the part of the researcher is required to populate the data repository.

Encouraging researchers to share their data comes with its own unique challenges, including clarifying what constitutes data, coping with the wide variety of data formats, fulfilling the need for highly descriptive metadata, overcoming fear of misuse and scooping, answering questions of ownership and rights, ensuring confidentiality of research subjects, providing access controls, and communicating effectively with a variety of unique research cultures, sometimes within the same department.<sup>4</sup>

Promotion of the data repository and data services in general requires clear expression of the benefits of data sharing to researchers. Sharing research data allows for the reproduction and verification of research, provides greater transparency by making publically funded research widely available, increases efficiency, encourages new uses for existing data, and speeds up the advancement of science.<sup>5</sup> However, an analysis by Borgman of the reasons why researchers share data and who benefits from data sharing showed that researchers are rarely the direct beneficiaries of sharing their data.<sup>6</sup> These findings suggest that to be successful, promotional strategies should focus on the importance of sharing data for the purposes of reproduction and verification, which can add to the researcher's reputation and aid in the advancement of science in their area of research. A study by Fecher asserted that data sharing requires the implementation of policies and incentives for researchers to share their data.<sup>7</sup> Furthermore, there are differences in attitude and motivational factors dependent upon the discipline or research culture.<sup>8</sup> Promotion and outreach strategies targeting the data sharing as a means of advancing knowledge in their field, by facilitating reproduction and verification, while advocating for the implementation of incentives for sharing data may be effective in increasing submissions to the repository.

Incorporating data management into the curriculum or offering data management workshops can grow understanding of the benefits of data sharing and data scholarship.<sup>9</sup> In particular, data management instruction to graduate students and researchers early in their careers, as they are first forming their research habits and workflows, creates awareness of data scholarship, including the benefits of data sharing and the importance of preserving and describing research data for future use.<sup>10</sup> Furthermore, addressing the data skills gap in librarians is a necessary step in developing research data services.<sup>11</sup> Data management workshops provide in-house opportunities for staff to develop data management skills and become familiar with the data needs of researchers and the data services offered by the institution.

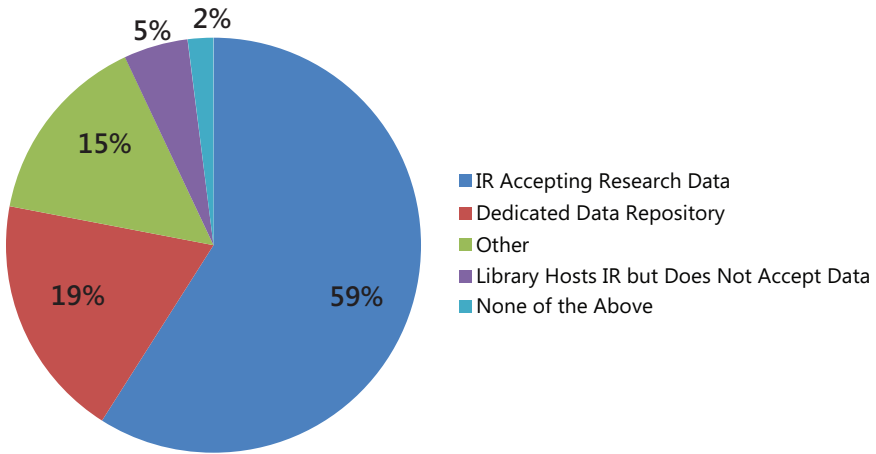
It is imperative that librarians and repository managers develop outreach and promotional strategies to increase awareness of the repository and association of the library with data services. According to a 2012 ACRL white paper, “a quarter to a third of all academic libraries are planning to offer some [research data services] within the next two years.”<sup>12</sup> The survey of 221 academic libraries showed 14.5 percent providing access to a repository or access and discovery services for data and almost 32 percent planning to offer these services in the next twenty-four months.<sup>13</sup> With so many libraries planning repository and data services, developing an understanding of the most effective outreach and marketing strategies for data repositories can save time and money.

Although repository infrastructure and organizational cultures differ, thus creating a need for varied promotional and outreach practices, commonalities can be found and adapted to unique institutional environments. To discover how data repositories are being promoted to their communities, an online survey and interviews were conducted with repository managers, data curators, and liaison librarians. This research identifies promotional techniques and key audiences and concludes with suggestions for developing marketing and outreach strategies.

## The Survey

In the summer of 2015, a brief online survey was distributed to the following library-focused research data lists in the United States: the Research Data, Access and Preservation (RDAP) e-mail list of the Association for Information Science and Technology; <http://mail.asis.org/mailman/listinfo/rdap>) and the American Research Libraries Data Sharing Support Google group (<https://groups.google.com/forum/?fromgroups#!aboutgroup/arl-data-sharing-support-group>). The survey remained open for two weeks. See appendix 9A for the full survey. Thirty-six people responded ( $n=36$ ). The small number of responses, coupled with the use of convenience sampling, prevents the generalization of the results to all institutions with data repositories and services.

Responses to the first question, “What type of data repository are you affiliated with?” indicated that the majority of responders offering data repository services are using their existing institutional repositories to store research data (59%) as shown in figure 9.1. The second question provided a multiple-choice list of outreach and promotion techniques. There were six choices that allowed respondents to indicate how they tailor their promotional techniques to the individual or department and seven choices with more generalized techniques. Generalized techniques were cited 127 times and tailored techniques 104 times. The most common form of outreach and promotion being used was recommending data repositories when asked (80%) as shown in table 9.1. The full data set including the original survey is published online.<sup>14</sup>



**FIGURE 9.1**  
What type of data repository are you affiliated with? (n=36)

**TABLE 9.1**  
Current Promotion Techniques for Data Repositories Online  
Survey Responses (n=36)

Classification	Promotional Technique	Number of Respondents
Tailored	Recommending data repositories when asked (e.g., reference question)	28
Tailored	Incorporate the repository into data management instruction	26
Generalized	Link to the repository on a subject page or LibGuide	25
Generalized	Link to the repository on the library's home page	23

**TABLE 9.1** (continued)

<b>Classification</b>	<b>Promotional Technique</b>	<b>Number of Respondents</b>
Generalized	Include the repository in boilerplate language for data management plans (DMPs)	22
Tailored	Targeted promotion to departments via liaisons and/or subject specialists	21
Generalized	Link to repository from data management tools on campus	20
Tailored	Individual e-mails to researchers	14
Generalized	Distribute paper-based promotional materials (e.g., flyers, brochures)	13
Generalized	Utilize institutional or library-run social media (e.g., Facebook, Twitter)	13
Generalized	Mass e-mails to the institutional community	11
Tailored	Incorporate the repository into information literacy instruction	10
Tailored	Technical solution that connects the repository to other campus research environments (e.g., storage, collaboration tools)	5
NA	Other	4
NA	None of the above	3

Question 3 provided respondents the opportunity to provide additional information about their approach. One respondent wrote, “We just launched [our repository] in January. Still trying to get folks to deposit things, including data. After a campus-wide marketing campaign, we are now going door-to-door via subject liaisons.” Another respondent said, “There is not great interest in promoting our IR for data as we can only take small datasets and it’s really only for fixed data—researchers are looking for a space that carries them throughout workflows.” Another respondent wrote, “We launched with a robust communications plan. In addition to an email by our University Librarian and our VP for Research to all faculty announcing the new service, we also had several presentations at prominent researcher meetings, including the Senate Research Committee, the meeting of the Associate VPs for Research (for each college) and a meeting of the department IT directors. This gave us exposure to a core group of faculty and we have seen a response that appears to be word-of-mouth distribution of the service.” This sampling of responses exemplifies how outreach and promotion of the repository fits within the context of repository development and organizational culture. The final question of the survey was used to recruit subjects for follow-up

phone interviews, and nearly half of the respondents expressed an interest (47%) in talking more.

## The Interviews

The interviews took place via phone and were structured to obtain detailed information on the current and future promotion of library-associated data repositories. Interviewees were e-mailed the questions prior to the interview. The questions asked were, in this order:

1. Tell me a little about how you promote the data repository.
2. Can you describe a promotional technique that has worked well?
3. Can you describe a promotional technique that has not worked well or has otherwise been abandoned?
4. Who are the primary audiences you promote the repository to?
5. How successful would you say your promotion of the data repository has been? How do you measure successful promotion of the data repository?
6. What has been the biggest challenge to increasing awareness of the data repository?
7. Do you promote the data repository differently than the institutional repository is promoted? How?
8. When thinking of ways to promote the data repository where do you look for inspiration?

Potential interviewees were initially contacted via e-mail. Of the seventeen who initially expressed interest in being contacted, fifteen responded and were successfully interviewed, and one interviewee was added who did not respond to the survey but was suggested by another interviewee ( $n=16$ ).

Included in the interviews are three representatives from repositories that have been accepting data for five to seven years, with nine participants representing repositories that have been accepting data for one to four years and four institutions not currently hosting data in their repositories but planning to host data or promoting disciplinary repositories to researchers in need of a data storage and access solution. Those affiliated with institutions not hosting data repositories were asked to answer the interview questions in relation to how they promote nonaffiliated repositories and data services to assist researchers in locating the best home for their data.

## *Measuring the Success of Repository Promotions*

Most of the interviewees stated they are not assessing the success of their promotional activities. Methods for tracking success mentioned by interviewees includ-



ed tracking workshop attendance and retrieval of metrics for mass e-mails, such as open and click-through rates. One interviewee stated, “Success is any time we aren’t dragging researchers to the repository.” Some interviewees indicated that an uptick in contact from researchers immediately following a promotional activity provides a means of gauging success. One interviewee cited the length of time between the promotional activity and when researcher workflows reach the stage of data deposit as presenting a significant challenge to measuring the success of the activity.

## *Successful Promotional Techniques*

Several promotion and outreach techniques were mentioned repeatedly across interviews. Forming partnerships with the sponsored programs office or the graduate school was discussed most often as an important and successful means of promotion. Interviewees cited a strong relationship with their sponsored programs office as a useful conduit between researchers applying for and receiving grants requiring data management plans that prefer researchers to make their data publically accessible. Likewise, responses indicated the role that grants administration staff play in alerting repository managers of researchers applying for grants who can then be contacted. One interviewee discussed how the research office suggests the university repository to researchers and even incorporates the repository into brochures and materials provided by the office. Several interviewees noted that as theses and dissertations are submitted, the graduate school can suggest the repository to graduate students as a storage and access solution for data related to their thesis and dissertation.

Face-to-face communication with key audiences one-on-one or in small groups was also repeated as a successful means of increasing awareness and use of the data repository. Personal communication with researchers was seen as highly beneficial particularly when conversations occurred at the point in the researcher’s workflow when a data storage and access solution is needed. Similarly, presentations to departmental administrators were reported as working well. According to one interviewee, “Presenting information on the repository and associated services in a way that catches the attention of administrators can cause a trickle-down effect where the information gets passed to researchers from trusted and authoritative sources.”

Some interviewees listed traditional promotional materials such as flyers, postcards, digital signage, and videos as successful techniques, while some interviewees found them to be ineffective. On the other hand, one interviewee pointed to the (poor) design of a flyer as the reason it did not have a high impact.

## *Unsuccessful Promotional Techniques*

Across the interviews, two techniques were repeatedly described as being unsuccessful: e-mail promotion and relying on liaisons to promote the repository to their liaison departments. Some interviewees cited using mass e-mail as a promotional technique because it is seen as a quick and easy way to hit a wide audience. One interviewee acknowledged that time and thought must be put into the composition and the logistics of sending the e-mail to maximize effectiveness. Another interviewee cited analytic data from an e-mail campaign. The data suggested that when the e-mail came from the library, fewer people opened it than when it originated from the office of the vice president for research.

Interviewees made some suggestions for supporting liaisons in the outreach role they play for the repository. First, the consensus was that liaisons needed a simple message to deliver. Providing them with talking points and working with them to construct “elevator speeches” can assist liaisons when presenting information on the repository to researchers and administrators. Second, respondents noted that librarians should also be included as a target audience for repository promotions. The more that library liaisons know about the repository and its services, the more confidently they can communicate the benefits of the repository to researchers. Third, interviewees noted the importance of advocates, and the need to make the repository a priority for library administrators.

## *Target Audiences*

All interviewees were based at academic research libraries. They all cited research and teaching faculty as their target audience for outreach and promotion of the data repository. Other audiences mentioned were small data researchers, graduate students, sponsored programs offices, departmental administrators, information technology staff, and research offices. One interviewee listed librarians as a target audience.

## *Challenges to Increasing Awareness*

Reaching researchers at the proper time in their workflow, when a data storage and access solution is needed, was the most cited challenge to the success of the repository. Other challenges to increasing awareness of the repository were countering the traditional view of the library, tailoring promotion and submission procedures to fit multiple types of data, overcoming researcher concerns over scooping, creating an awareness of the scholarship of data, and gaining full support of the administration in order to make the repository an organizational or departmental priority.

## *Differences in Promoting the Institutional Repository and the Data Repository*

Institutional repositories are better-established than data repositories, as is the scholarly communication system around the publication of research findings in the form of text. In many cases, interviewees stated they promote the institutional repository and data repository together. One interviewee stated they have been promoting them together but would like to see a separation of the two. One interviewee noted the exceptional reputation of the institutional repository and stated, “We ride on their coattails when we can.” Some interviewees noted the different benefits to the researcher of the institutional repository (impact and recognition) and data repository (verification, reuse) as making it difficult to tie them together. One interviewee stated, “The institutional repository is promoted as a place to put one document at a time, the data repository is a place to store a body of work.” Another interviewee mentioned the fact that IR contents are easier to come by since work can be harvested from the web and theses and dissertation digitization projects and deposit mandates can help to populate it. In contrast, data repositories rely solely upon author submissions. As one interviewee noted, “It is early days yet to gauge the impact of funder and publisher data sharing requirements.”

## *Looking for Inspiration*

The most cited sources of inspiration for promotion and outreach techniques were conferences and looking to other institutional, disciplinary, and data repositories. Brainstorming with colleagues, marketing literature, and Twitter were also cited by interviewees. Several interviewees stated they haven’t yet thought much about promotion of the data repository out of concern for scalability of curation procedures and intake processes.

## **Discussion**

Overwhelmingly, promotion techniques cited in the interviews as successful for data repositories involved forming strategic partnerships with campus departments and targeting promotion to specific individuals or group. From this study it is impossible to say whether the perception of success is due to the immediate feedback received from targeted promotions or if targeted techniques increase use of the repository. Interestingly, in the initial survey generalized forms of pro-

motion (127) were cited more often than tailored forms (104). The most cited unsuccessful technique for promoting the repository was mass e-mail, cited as a promotional technique (11 times) in the initial survey. As researchers are inundated with e-mail, they may use sender addresses and subject lines to decide what to open. Generalized techniques such as e-mail, links, and boilerplate language for DMPs increase awareness of the repository and serve the necessary function of providing information to those actively seeking information. Without analytic data it is not possible to definitively ascertain the effectiveness of these methods. However, interviewees' overall perceptions of the success of promotional techniques they have used can provide useful insight.

Not surprisingly, targeted promotion of the repository and data services to key stakeholders (researchers, sponsored programs offices, and graduate schools) is viewed as a successful means of raising awareness of the repository. As researchers interact with sponsored programs offices and graduate schools at key points in the research process, the staff in these offices are in a position to direct people to the repository at the time of data-sharing need. Embedding the repository in the promotional materials and workflow of sponsored programs and graduate school offices is one way to overcome the challenge of timing to reach researchers at the appropriate stage in the research process.

Librarians and library staff must be remembered as a key stakeholder group to be targeted for outreach and promotion activities. The perceived lack of success of liaison librarians in promoting the repository to their respective departments may be due to the need for an improved understanding of data services and data sharing. When promoting and designing data management and data services workshops, include liaisons and reference librarians as a target audience. Workshops can provide a means for librarians to develop their skills with data and gain an understanding of data services provided by the institution. Successfully leveraging the relationships that liaisons have with researchers and departments may provide a vital boost to awareness of the repository and data submissions.

The promotion challenges are different for the data repository and the institutional repository, yet they have much in common. When fighting the uphill battle toward wide adoption of data services, it can be helpful to look to successful promotion techniques used by the institutional repository for guidance and support. Additionally, while the workflows for the IR and data repository differ remarkably, consideration should be given to presenting them together as parts of a distributed institutional solution for disseminating and preserving research products. Indeed, marketing the IR and data repository together may aid in legitimizing library stewardship of research products in the eyes of stakeholders.

When planning future repository outreach and promotion, here are some key lessons learned from this study:

*Identify Partners to Reach Your Target Audiences:* Partnerships with the graduate school, information technology, and sponsored programs offices are inval-

able in raising awareness of the repository and data services. Consider: Who are other potential stakeholders? What departments or researchers have pressing data-sharing needs? What is the optimal granularity when identifying target audiences (individual colleges, disciplines, research groups)?

*Identify Incentives and Benefits for Researchers:* Taking stock of the potential benefits to researchers will provide talking points and aid in the creation of the web presence, e-mails, and other promotional materials. Consider: Do the benefits and incentives differ for different user groups? What are the benefits that remain constant across groups?

*Develop Learning Opportunities:* Workshops not only provide an opportunity to instruct students and researchers on data management, services and the repository, but they also provide professional development opportunities for librarians as well. Consider hosting a workshop targeted to library faculty and staff, and seek out graduate or undergraduate courses in which to offer data management and data literacy as one-shot workshops.

*Consider Marketing the IR and Data Repository a One-Stop Solution:* The IR and data repository may internally be viewed separately, but pairing them for marketing reasons presents a more comprehensive solution for researcher dissemination and preservation needs. Consider: Is it beneficial to brand them as one or separately? If you brand them separately, can you market them together? Does choosing to promote them together broaden the target audience? Are there other institutional tools or systems that can be linked for marketing purposes?

*Get Feedback:* As with any service or initiative, gaining feedback from stakeholders will provide information crucial to developing user-driven services. Feedback will help identify the perceived benefits from the user's perspective that can be incorporated into future promotion and outreach planning. Consider how you will assess the impact of your promotional activities.

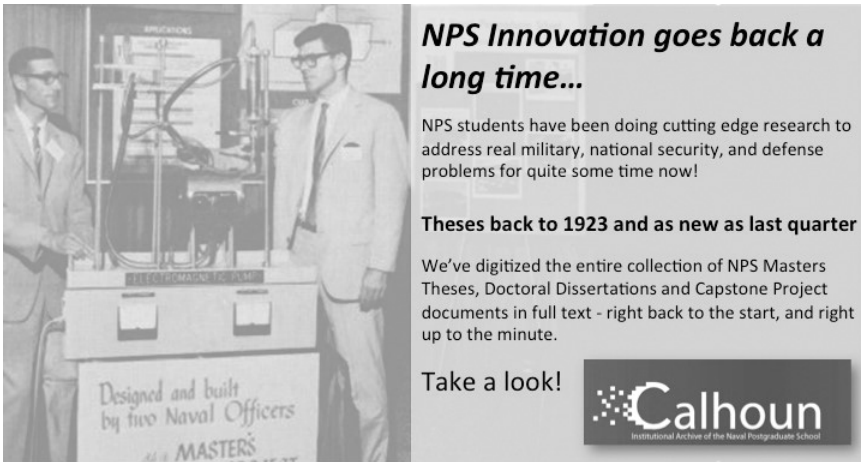
## Conclusion

Aside from a very few examples, academic library-run repositories for digital data are a new development. Data scholarship is not widely understood and the terms *library* and *data* are not synonymous. In these early days, promoting the data repository is no simple task. It is not a one-person job. It requires building upon the library's reputation for outreach, preservation, curation, and understanding of the scholarly communication system, then linking those skills and that knowledge to research data in the minds of your target audiences. Repository managers need to build and strengthen partnerships with repository stakeholders, educate current and future researchers on the benefits of depositing their data into a repository, and establish their data repositories as a key part of the emerging infrastructure for scholarly communication.

# Promotional Examples for Inspiration

In the list below and figures 9.2, 9.3, and 9.4 are some examples of successful promotional materials for data repositories and data services:

- Data Repository for the University of Minnesota e-mail sent to all faculty: <http://continuum.umn.edu/email/2015/drum/>.
- Library Guide for Research Data Management at Virginia Commonwealth University: <http://guides.library.vcu.edu/data>.
- Scholar@UC Press Kit from the University of Cincinnati: <http://libapps.libraries.uc.edu/scholarblog/scholaruc-press-kit/>.
- Office of the Vice President of Research at Purdue University brochure (pp. 8–9): <https://www.purdue.edu/research/docs/pdf/Winter2013.pdf>.
- Purdue University Research Repository You Tube video: <https://youtu.be/Yw0IJj7FqA8>.
- University of New Mexico, Digital Data Management, Curation, and Archiving Library Guide: <http://libguides.unm.edu/data>.



***NPS Innovation goes back a long time...***

NPS students have been doing cutting edge research to address real military, national security, and defense problems for quite some time now!

**Theses back to 1923 and as new as last quarter**

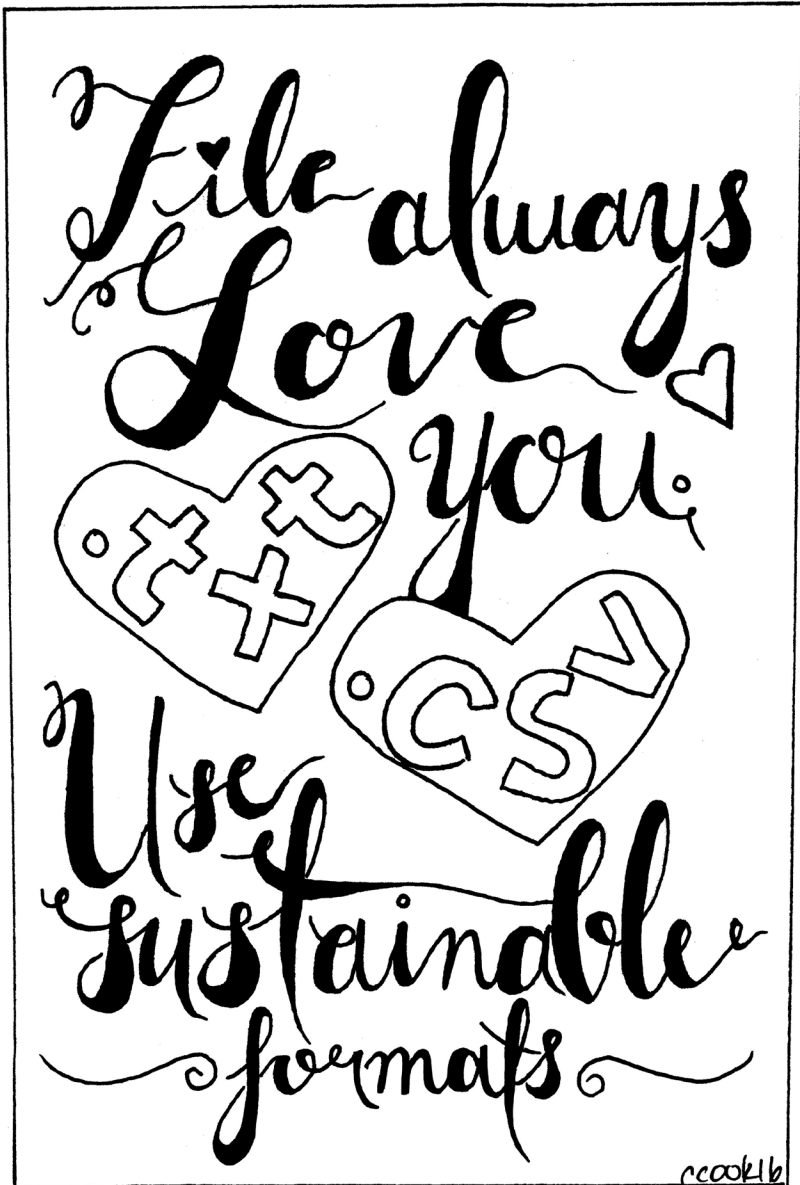
We've digitized the entire collection of NPS Masters Theses, Doctoral Dissertations and Capstone Project documents in full text - right back to the start, and right up to the minute.

Take a look!

**Calhoun**  
Institutional Archive of the Naval Postgraduate School

**FIGURE 9.2**

Slide promoting institutional and data repository at US Naval Postgraduate School, Irene Berry. This image is not eligible for copyright protection in the US.



Research Data Services | [researchdata.wisc.edu](http://researchdata.wisc.edu) | [@UWMadRschSvcs](https://twitter.com/UWMadRschSvcs)

**FIGURE 9.3**

University of Wisconsin–Madison Data Comic by Cameron Cook, CC-BY-SA 2.0. Comic posted on Twitter February 2016.



Research Data Services  
UNIVERSITY OF WISCONSIN-MADISON

## Data got you down? We can help!

### data management plans

We'll work with you to draft a plan for your grant application.

### personal consultations

Have a data management question? We're here to help.

### education & training

Access our educational materials & work with us to plan a training for your lab or department.

```
01000    10101
0001000100 0011110001
10001010101010010100101010
0000010000100100001011001010
0010101001010010100101010010
01001010101010101010101010
0110101010010101001010
01101010101001010
10110101001
010110
10
```



We love your data too.

[researchdata.wisc.edu](http://researchdata.wisc.edu) & [@UWMadRschSvcs](https://twitter.com/UWMadRschSvcs)

### FIGURE 9.4

University of Wisconsin–Madison Research Data Services flyer by Brianna Marshall, CC BY 2.0.



# Acknowledgments

I would like to acknowledge and thank the following people for their assistance in the writing of this chapter: Lisa Johnston at the University of Minnesota, Margaret Henderson at Virginia Commonwealth University, Michael Witt at Purdue University, Abigail Goben at the University of Illinois–Chicago, Brian Westra and Catherine Flynn-Purvis at the University of Oregon, Ann Nurnberger at Columbia University, Amy Koshoffer at University of Cincinnati, David Fearon at Johns Hopkins University, Chris Eaker at the University of Tennessee–Knoxville, John Wheeler at the University of New Mexico, Holly Miller at the Florida Institute of Technology, Irene Berry at the Naval Postgraduate School, Sara Mannheimer at the University of Montana, Amy Buckland at the University of Chicago, Brianna Marshall at the University of Wisconsin-Madison, and Christina Chan-Park at Baylor University.

# Appendix 9A: Data Repository Promotional Practices—Initial Google Survey

Please take a moment to complete this brief (1–2 minute) survey to discover how academic libraries are promoting their data repositories on campus. The survey will close on Tuesday, July 28, 2015, and the de-identified results will be openly shared with the listserv shortly thereafter.

## What kind of repository are you affiliated with? (Choose all that apply)

- Our library hosts a dedicated data repository.
- Our library hosts an institutional repository that accepts research data.
- A campus unit outside the library offers a data repository service.
- Our library hosts an institutional repository but it does not accept data.
- None of the above

## How do you currently promote the data repository? (Choose all that apply)

- Individual emails to researchers
- Mass emails to the institutional community
- Targeted promotion to departments via liaisons and/or subject specialists
- Incorporate the repository into data management instruction
- Incorporate the repository into information literacy instruction
- Distribute paper-based promotional materials (e.g. flyers, brochures)
- Utilize institutional or library run social media (e.g. facebook, twitter)
- Link to the repository on the library's home page
- Link to the repository on a subject page or LibGuide
- Link to repository from data management tools on campus
- Include the repository in boilerplate language for data management plans (DMP's)
- Recommending data repositories when asked (eg. reference question)
- Technical solution that connects the repository to other campus research environments (e.g. storage, collaboration tools)
- None of the above

## Tell us more about your approach.

We would appreciate the name of your institution or data repository in your answer.

**If we may contact you for a brief follow-up interview, please include your name and email.**

# Notes

1. Richard K. Johnson, "Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication," *D-Lib Magazine* 8, no. 11 (November 2002), <http://www.dlib.org/dlib/november02/johnson/11johnson.html>; Philip M. Davis and Matthew J. L. Connolly. "Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace." *D-Lib Magazine* 13, no. 3–4 (2007), <http://dialnet.unirioja.es/servlet/articulo?codigo=2284108>.
2. G. Sayeed Choudhury, "Case Study in Data Curation at Johns Hopkins University," *Library Trends* 57, no. 2 (2008): 211–20, doi:10.1353/lib.0.0028.
3. Ibid.; Michael Witt, "Institutional Repositories and Research Data Curation in a Distributed Environment," *Library Trends* 57, no. 2 (2008): 191–201, doi:10.1353/lib.0.0029; Wong, Gabrielle K. W. Wong, "Exploring Research Data Hosting at the HKUST Institutional Repository," *Serials Review* 35, no. 3 (2009): 125–32, doi:10.1080/00987913.2009.10765229.
4. Christine L. Borgman, "The Conundrum of Sharing Research Data," *Journal of the American Society for Information Science and Technology* 63, no. 6 (2012): 1059–78, doi:10.1002/asi.22634; Benedikt Fecher, Sascha Friesike, and Marcel Hebing, "What Drives Academic Data Sharing?" *PloS One* 10, no. 2 (2015): e0118053, doi:10.1371/journal.pone.0118053; Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame, "Data Sharing by Scientists: Practices and Perceptions." *PloS One* 6, no. 6 (2011), Public Library of Science: e21101, doi:10.1371/journal.pone.0021101.
5. Borgman, "The Conundrum of Sharing Research Data."
6. Ibid.
7. Fecher, Friesike, and Hebing, "What Drives Academic Data Sharing?"
8. Katherine G. Akers and Jennifer Doty. "Disciplinary Differences in Faculty Research Data Management Practices and Perspectives," *International Journal of Digital Curation* 8, no. 2 (2013): 5–26, doi:10.2218/ijdc.v8i2.263; Lisa M. Federer, Ya-Ling Lu, Douglas J. Joubert, Judith Welsh, and Barbara Brandys, "Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff," *PloS ONE* 10, no. 6 (2015): e0129506, doi:10.1371/journal.pone.0129506; Youngseek Kim and Melissa Adler, "Social Scientists' Data Sharing Behaviors: Investigating the Roles of Individual Motivations, Institutional Pressures, and Data Repositories," *International Journal of Information Management* 35, no. 4 (August 2015): 408–18; Youngseek Kim, and C. Sean Burns, "Norms of Data Sharing in Biological Sciences: The Roles of Metadata, Data Repository, and Journal and Funding Requirements," *Journal of Information Science* 42 (April 2016): 230–45.
9. Melissa H. Cragin, Carole L. Palmer, Jacob R. Carlson, and Michael Witt, "Data Sharing, Small Science and Institutional Repositories." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 368, no. 1926 (2010): 4023–38, doi:10.1098/rsta.2010.0165; Megan Sapp Nelson, "Data Management Outreach to Junior Faculty Members: A Case Study," *Journal of eScience Librarianship* 4, no. 1 (2015): e1076, doi:10.7191/jeslib.2015.1076; Don MacMillan, "Data Sharing and Discovery: What Librarians Need to Know," *Journal of Academic Librarianship* 40, no. 5 (September 2014): 541–49, doi:10.1016/j.acalib.2014.06.011.

10. Cragin et al., “Data Sharing, Small Science and Institutional Repositories”; Sapp Nelson, “Data Management Outreach to Junior Faculty Members.”
11. Sheila Corral, Mary Anne Kennan, and Waseem Afzal, “Bibliometrics and Research Data Management Services: Emerging Trends in Library Support for Research,” *Library Trends* 61, no. 3 (2013): 636–74, doi:10.1353/lib.2013.0005.
12. Tenopir, Birch, and Allard, *Academic Libraries and Research Data Services*, 3.
13. *Ibid.*, 19.
14. Katherine Gerwig and Lisa R. Johnston, “Promotional Practices of Research Data Repositories,” dataset, 2015, doi:10.13020/D6J01R.

## Bibliography

- Akers, Katherine G., and Jennifer Doty. “Disciplinary Differences in Faculty Research Data Management Practices and Perspectives.” *International Journal of Digital Curation* 8, no. 2 (2013): 5–26. doi:10.2218/ijdc.v8i2.263.
- Borgman, Christine L. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63, no. 6 (2012): 1059–78. doi:10.1002/asi.22634.
- Choudhury, G. Sayeed. “Case Study in Data Curation at Johns Hopkins University.” *Library Trends* 57, no. 2 (2008): 211–20. doi:10.1353/lib.0.0028.
- Corral, Sheila, Mary Anne Kennan, and Waseem Afzal. “Bibliometrics and Research Data Management Services: Emerging Trends in Library Support for Research.” *Library Trends* 61, no. 3 (2013): 636–74. doi:10.1353/lib.2013.0005.
- Cragin, Melissa H., Carole L. Palmer, Jacob R. Carlson, and Michael Witt. “Data Sharing, Small Science and Institutional Repositories.” *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 368, no. 1926 (2010): 4023–38. doi:10.1098/rsta.2010.0165.
- Davis, Philip M., and Matthew J. L. Connolly. “Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University’s Installation of DSpace.” *D-Lib Magazine* 13, no. 3–4 (2007). <http://dialnet.unirioja.es/servlet/articulo?codigo=2284108>.
- Fecher, Benedikt, Sascha Friesike, and Marcel Hebing. “What Drives Academic Data Sharing?” *PLoS One* 10, no 2 (2015): e0118053. doi:10.1371/journal.pone.0118053.
- Federer, Lisa M., Ya-Ling Lu, Douglas J. Joubert, Judith Welsh, and Barbara Brandys. “Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff.” *PLoS ONE* 10, no. 6 (2015): e0129506. doi:10.1371/journal.pone.0129506.
- Gerwig, Katherine and Lisa R. Johnston. “Promotional Practices of Research Data Repositories.” Dataset. 2015. doi:10.13020/D6J01R.
- Johnson, Richard K. “Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication.” *D-Lib Magazine* 8, no. 11 (November 2002). <http://www.dlib.org/dlib/november02/johnson/11johnson.html>.
- Kim, Youngseek, and Melissa Adler. “Social Scientists’ Data Sharing Behaviors: Investigating the Roles of Individual Motivations, Institutional Pressures, and Data Repositories.” *International Journal of Information Management* 35, no. 4 (August 2015): 408–18.

- Kim, Youngseek, and C. Sean Burns. "Norms of Data Sharing in Biological Sciences: The Roles of Metadata, Data Repository, and Journal and Funding Requirements." *Journal of Information Science* 42 (April 2016): 230–45.
- MacMillan, Don. "Data Sharing and Discovery: What Librarians Need to Know." *Journal of Academic Librarianship* 40, no. 5 (September 2014): 541–49. doi:10.1016/j.acalib.2014.06.011.
- Sapp Nelson, Megan. "Data Management Outreach to Junior Faculty Members: A Case Study." *Journal of eScience Librarianship* 4, no. 1 (2015): e1076. doi:10.7191/jeslib.2015.1076.
- Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. "Data Sharing by Scientists: Practices and Perceptions." *PloS One* 6, no. 6 (2011). Public Library of Science: e21101. doi:10.1371/journal.pone.0021101.
- Tenopir, Carol, Ben Birch, and Suzie Allard. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future*. ACRL white paper. Chicago: Association of College and Research Libraries, 2012. [http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir\\_Birch\\_Allard.pdf](http://www.ala.org/acrl/sites/ala.org/acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf).
- Witt, Michael. "Institutional Repositories and Research Data Curation in a Distributed Environment." *Library Trends* 57, no. 2 (2008): 191–201. doi:10.1353/lib.0.0029.
- Wong, Gabrielle K. W. "Exploring Research Data Hosting at the HKUST Institutional Repository." *Serials Review* 35, no. 3 (2009): 125–32. doi:10.1080/00987913.2009.10765229.





**PART III**  
**Preparing Data for  
the Future**

*Ethical and Appropriate  
Reuse of Data*







## CHAPTER 10\*

# Open Exit

## Reaching the End of the Data Life Cycle

*Andrea Ogier, Natsuko Nicholls, and Ryan Speer*

### Introduction

Scientific research data often has a longer lifespan than the project that creates it. This is particularly true when good research data management practice is put into place. Good data management throughout the data life cycle is essential for successful long-term preservation and sharing, ensuring a long life span of use for research data. In addition to good data management, many of us would agree that the importance, impact, and relevance of one's research data often influences the potential long-term value of it—that is that relevance, value assessment, and retention are all closely linked. Yet it remains uncertain whether or not the retention of data increases its inherent value. More fundamentally, does data in the life cycle smoothly progress from one stage to another without a gap or an exit? Should review, assessment, and evaluation functions for scientific records and data be included at every stage prior to reaching the end of the data life cycle? These questions and similar inquiries about the life span (and “death” of data<sup>1</sup>) have motivated us to investigate a variety of actions involved in curation decisions for data retention or deletion.

In this chapter, we suggest that potential use or retention should be considered by researchers and data curators in every phase of the data life cycle,

---

\* This work is licensed under a Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

particularly in the life cycle “potholes” where the cycle could naturally slow, stall, or end. We argue that identifying and preparing for these points is a vital part of data curation in which long-term value is of central importance. From a curation practitioner’s perspective, value assessment of all kinds of records, including data sets, is a crucial part of appraisal and selection for records management, curation, and collection development. These appraisal and selection activities are the iterative, responsive, and active processes of reappraisal, weeding, deselection, deaccession, and disposition. These actions are backed up by a variety of technical, legal, and institutional policies. And appraisal activities should occur throughout the research process and data life cycle and should be based on criteria rather than on the assumption that the very act of long-term preservation implies value.

In order to advance our understanding of the actions and decisions that adequately safeguard data for future use, we examine a variety of technical, legal, and institutional responses, controls, and resources that influence actions (and the actors involved in these actions) to retain or not retain the data. Three areas provide context for discussion: university records and information management, library collections management, and data curation. University records and information management, hereafter “records management,” has grown out of a concern for records as corporate assets that must be managed according to a specific set of practices set by a local regulatory environment.<sup>2</sup> Similarly, library collections management (or “collection development”) is understood as a set of routines aimed at adding materials, removing materials, and efficiently finding materials in a library’s collection. We believe this comparative exploration, bringing the discourse and practices developed by well-articulated records management and library collections philosophies alongside the formative practices of data curation, will help us identify points in the data life cycle where curation would (or should) come to an end.

## Comparative Exploration

In her article exploring a selection and appraisal framework for digital curation, Jinfang Niu adopted a comparative approach based on the processes and theories from the archives and records management communities.<sup>3</sup> We take a similar approach; however, as Niu draws from methodologies aimed solely at selecting digital objects for preservation, we broaden our focus by exploring methodologies aimed at deletion, disposition, and rejection of materials that exist as part of a collection. The distinction is slight, but important; we want to shed light on the diverse interpretations and understandings of how data should progress throughout the life cycle.

As we approach disposition and end-of-life-cycle issues from the three perspectives (university records and information management, library collections, and data curation), we focus on the following five areas:

1. terminology (usage and interpretation);

2. scope (types, formats, and uses of objects);
3. authority (actors and directives);
4. appraisal criteria (actions and factors that influence those actions); and,
5. resources (human, financial, and physical space).

Although some existing studies suggest as many as ten criteria for disposition (as it appears in routines of selection and appraisal),<sup>4</sup> we focus on these five elements not as criteria themselves, but as a basis for comparison in order to determine how items are excluded or removed from collections and archives. Tables 10.1 through 10.5, following a brief discussion, will showcase our comparative observation across the three areas.

## “End of Life Cycle” Terminology

Beginning with terminology allows us to draw out conceptual similarities and differences across the three areas to get a better sense of accepted definitions. As shown in table 10.1, the term *disposition*, which is a key term in records management, refers to a strictly bounded and regularly scheduled decision-making process where an item is either archived or destroyed.<sup>5</sup> The term *weeding* used in library collection management, for example, creates the mental image of a gardener removing weeds so that the carefully planted seeds can get more sunlight and rain, aligning these decisions with natural processes.<sup>6</sup> *Selection* and *deselection* link the additive and subtractive collections decisions, just as using *appraisal* and *reappraisal* creates a cyclic decision narrative in the realm of data curation. In this chapter we use the terminology native to the discipline considered in order to tie it more closely to the source material.

**TABLE 10.1**  
Comparison in End-Of-Life-Cycle Terminology

	<b>University Records and Information Management</b>	<b>Library Collection Management</b>	<b>Data Curation</b>
<b>Terminology</b>	Official record Active/inactive records Disposition: retention or destruction	Collection Maintenance Weeding Deaccessioning “Data-driven” deaccession Deselection	Digital content Retention Appraisal/reappraisal; Selection/acquisition Data transfer/ migration Disposition Destruction

**TABLE 10.2**  
Comparison in Scope

	<b>University Records and Information Management</b>	<b>Library Collection Management</b>	<b>Data Curation</b>
<b>Scope</b>	Theoretically embraces all information created by organization; includes any information created in support of the organization's mission or in fulfillment of its legal obligations.	Everything the library or archive subscribes to or collects, including provisions for gift and legacy materials: books, journals, digital resources, media, hardware, software, etc.	Theoretically everything that researchers generate out of research projects—recorded factual material commonly accepted in the scientific community as necessary to validate research findings. <sup>a</sup>
	Disposition of paper records is often more effective than electronic records management.	There is a difference between discarding the object and discarding the metadata.	Decisions are often influenced by types of data, state of data (e.g., raw, primary, analyzed, published) and the sensitivity of data.

a. OMB Circular A-110 defines data as “the recorded factual material commonly accepted in the scientific community as necessary to validate researching findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues” and has been widely adopted by many federal funding agencies. (Office of Management and Budget, “Uniform Administrative Requirements for Grants and Agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations,” OMB Circular A-110, November 6, 1999, [https://www.whitehouse.gov/omb/fedreg\\_a110-finalnotice](https://www.whitehouse.gov/omb/fedreg_a110-finalnotice).)

## Scope

The second element refers to the types, influence, and use of objects (whether physical, digital, or combined). Within the domain of records management, library collections, and data curation, we have explored the extent to which disposition decisions (or the lack thereof) about digital objects are based on the methodologies developed for physical objects as shown in table 10.2. A comparison of scope across these three areas demonstrates important distinctions between ideal scope and the reality of implementation. For example, records management encompasses all the documentation generated by an organization, but various factors, such as local policies and confidence in the disposition of secure documents, affect the ability of a records program to manage secure digital records in the same way as secure paper records. Library collections management distinguishes between the object (either

physical or digital) and the metadata representing that object; discarding the object and discarding the metadata in the library catalog are often two entirely separate processes. For data curation this distinction between ideal scope and factors that limit implementation may also be important where raw data contains sensitive information or is too large to be easily stored. In these situations the metadata may be an important representation for the data itself. What is interesting about data curation is that its scope is expanding at record speed, given the diverse formats of data and types of digital content that includes even research project websites, audio and video files, and geospatial information systems.<sup>7</sup> By contrast, some institutions like the National Oceanic and Atmospheric Administration (NOAA) narrowly define specific categories of scientific records (original data, synthesized products, and experimental products) as subject to the appraisal and disposition procedure.<sup>8</sup>

**TABLE 10.3**  
Comparison in Authority

	<b>University Records and Information Management</b>	<b>Library Collection Management</b>	<b>Data Curation</b>
<b>Authority (actors and directives)</b>	Records schedule Retention schedule Records manager or records coordinator Legal directives	Collection developer/manager Librarians State-level directives Consortial policies	Data steward Data producer Repository collection policies Institutional regulations Funder directives

## *Authority*

The third element broadly covers actors who have some control or power over decisions made about the object (see table 10.3). These actors may be in the form of people or positions within a larger organization (records manager, collections librarian) or in the form of policies, mandates, or laws. As an example, distinct characteristics of traditional records and information management approaches frame the death of data as more dependent on human factors than on the analysis of legal requirements. Where official records retention schedules are incomplete, long-term records appraisal must rely on professional judgment. In the world of research data management and curation, where policies are still being formed, this acknowledgement of human decisions above legal agency could legitimize evidence-based data exit strategies.

## Appraisal Criteria

The fourth element examines processes enacted upon the object; in this case, the decision to remove an object from a collection or life cycle. As shown in table 10.4, this decision-making process is rather highly developed in records management, which relies on extensive records retention schedules, comprehensive guidance created to identify disposition dates, and instructions for all types and categories of organizational information. Records schedules have admirable specificity, but schedule creators generally privilege administrative need and organizational legal obligations, which might only obliquely apply to the more uncertain environment of research data retention.<sup>9</sup>

Library collections weeding schedules are often marked by a concern for resources; as physical or digital space or budgetary resources become scarce, weeding projects are initiated and driven by a variety of criteria. In libraries where space and cost may not be critical issues, weeding projects can be driven by a concern for the “health” of the collection or a desire for managing the currency of the information.<sup>10</sup> Appraisal in data curation has developed to ensure that scientific records and data are usable over time; thus metrics of cost and historical use may not be entirely relevant. Perhaps the most urgent criterion for assessment in data curation is that of compliance; data that contains sensitive information, whether due to personally identifiable information or representing a security risk (e.g., credit card information), should be managed and disposed with a high degree of care.

**TABLE 10.4**  
Comparison in Appraisal Criteria

	<b>University Records and Information Management</b>	<b>Library Collection Management</b>	<b>Data Curation</b>
<b>Appraisal Criteria</b>	Criteria include liability administrative need superseded, obsolete, rescinded; time period after event/action.	Criteria include space currency subject coverage usage/cost-per-use duplication (in format or consortial location)	Criteria include funder ROI compliance (repository) collection alignment scientific/historical/continuing value of data (in terms of reusability) quality integrity

## Resources (Human, Financial, and Spatial)

The fifth element addresses the cost needed to maintain the object within the collection (see table 10.5). Apart from large paper records storage operations, records management can be a cost-effective force multiplier for data management: records managers are unique within organizations in that they are responsible for the disposition of information created by others. Libraries may find themselves grappling with a variety of concerns, including the cost of purchasing or licensing collections, the high value of library real estate (location in city or on campus, stacks vs. study space), or the quality of the metadata provided by vendors (where costly staff time may be needed). Like the records management or library collections areas, there is significant cost associated with data stewardship; however, the cost of data curation is still unknown. Recent studies and tools have emerged in Europe from the “Collaboration to Clarify the Costs of Curation” Project (also known as 4C), which aims to emphasize the value of investing in curation infrastructure.<sup>11</sup>

**TABLE 10.5**  
Comparison in Resources

	<b>University Records and Information Management</b>	<b>Library Collection Management</b>	<b>Data Curation</b>
<b>Resources</b>	Costs of staffing and centralized records management program; costs of staff time and resources for records management tasks within records-creating units	Costs: budget and subscription/purchase models, staffing resources, space resources (physical) digital space counted by numbers of titles/items rather than by storage size	Storage/backup costs Preservation costs Cost of creating and managing preservation metadata (to ensure discoverability)

## Discussion

By focusing on these five elements (terminology, scope, authority, actions/appraisal criteria, and resources), we now summarize processes in use across records management, library collections, and data curation in order to provide insight into practices of planned data retention and deletion.

## *University Records and Information Management*

In the discipline of records management, appraisal for records retention predominantly is concerned with the primary administrative use of information by the creating organization, with a general emphasis on addressing liabilities or inefficiencies associated with ongoing maintenance of the documents by the original creators. The secondary value of information, or the measure of its enduring utility for audiences outside of the creating unit or organization, is also a focus of records retention scheduling. However, primary use is often the first concern, and appraisal approaches associated with records management are notable for relying on authorities more familiar and significant to research administrators than to academic departments or information managers outside of the records profession.<sup>12</sup> A retention decision from the records realm will rely on formal legal requirements for record keeping (when available) and other guidance found in state records retention guidance (when applicable and present), federal statute and administrative law, or on local institutional (e.g., university) policies based on business needs.<sup>13</sup>

At its heart, records management is centered upon the idea of the “record,” which may be deemed “official” as the product of state business or governance, “active” in that it is considered current, or “inactive.” These categories can affect the retention schedule and disposal method along with the content or coverage of the record. Thinking about research data as an official record can introduce novel approaches to determining how to retain and dispose of data, potentially offering new perspectives from which to address some problematic situations in data curation. For example, considering your local records management approaches to sensitive or confidential data sets may be informative to those developing data management plans or data retention policies; though the majority of research data may not be governed by an externally mandated retention/disposal schedule, the data could fall under the mandate of other local policies intended to govern information access and security, such as those maintained by institutional review boards or related administrative units. Consulting local or state-level records management policies regarding issues of liability and security could help in answering questions about data retention and inform the decision to remove a data set from the curatorial process.

## *Library Collections*

Creating policies and identifying the criteria for removing items from a collection has long been a part of maintaining a healthy library collection.<sup>14</sup> Library collections have grown beyond just the physical; however, in many libraries, phys-



ical volume counts still function as a metric for library status,<sup>15</sup> and if appraisal criteria for digital collections exist, they are often based on the same criteria for print-based collections. Library collection processes have identified a variety of criteria for what physical materials to withdraw (or weed) such as appearance, duplication in other collections, outdated content, and low usage.<sup>16</sup> In most libraries, where space is often at a premium, physical dimensions and shelf space are also a concern; the ever-expanding suite of library services necessitates careful consideration of physical collections.

Like records management, library collections concern both physical and electronic records; however, a library's digital collections, such as e-books, e-journals, and other e-resources, present slightly different concerns. While currency remains an issue for electronic materials, as Mike Waugh and colleagues noted in their discussion of an e-book weeding project conducted at LSU,<sup>17</sup> concerns over physical space and appearance do not apply to digital collections; the e-book weeding project at LSU was based on criteria of currency rather than space. However, physical concerns could easily translate to criteria of financial resources or cost: digital collections are usually hosted by the publisher and provided to libraries on a subscription model. While they don't require physical space within the library, the monetary cost of these resources could be a critical factor for retention or deselection. Metrics of cost-per-use are emerging as vitally important criteria for assessing digital resources and are figuring into deselection policies and activities, though they are not without significant drawbacks.<sup>18</sup>

In addition to concerns over space and cost, criteria for deselection may also be set by membership in consortia or agreements with multi-institutional digital libraries. HathiTrust, for example, uses member institutions' print holdings to determine legal use of in-copyright digital materials; in order for a user at a member institution to gain access to the digital copy of an in-copyright work, their institution must have at one time owned a print copy of the work.<sup>19</sup> In this scenario, a physical volume could be removed from the collection without losing access to the digitized copy; however, it may be resource-intensive to do so, and special care must be taken to ensure that the correct metadata record for the digital copy remains. Similarly, membership in state or regional library consortia may affect these decisions. For example, the Association of Southeastern Research Libraries (ASERL) has formed a cooperative print journal retention policy and joined the Washington Research Library Consortium in forming a print journal archive.<sup>20</sup> In these agreements, libraries agree to retain certain print materials for a specified amount of time (in the case of ASERL, until 2035). Thus, these consortial agreements and memberships influence what can and cannot be removed from the collection.

While considering data as "just another" library collection may gloss over some of the uniqueness that emerged from the disciplines of data curation and data management, it also presents a history of suitable criteria that could be used

to assess research data. The term *data curation* itself implies a curatorial framework of management; merely uploading data into a digital or institutional repository is not data curation, nor is it good data management. These collection management processes and criteria, used for decades by librarians to curate and care for physical (and now digital) library collections, could serve as an initial framework for assessing whether data should stay or exit the research life cycle.

## Data Curation

While records management and library collections practices could inform deselection within the data life cycle, often data retention is assessed only at the end of a project when the researcher, often attempting to comply with funder sharing requirements, determines which data to deposit into an archive. In these cases, the appraisal and deselection practices within records management and library collections are applied, but only after the object is in its final form. Digital curation, which is defined as “maintaining and adding value to a trusted body of digital information for future and current use; specifically, the active management and appraisal of data over the entire lifecycle,”<sup>21</sup> needs to operate in situ: before, during, and after the research process. Unfortunately, representing the research process in a life cycle implies that the transition is seamless and smoothly progress from stage to stage. However, Carlson argues, “the most critical gap between the stages in a life cycle model is between the stages where the data are actively managed for use by the researcher who developed the data to where the data transition into being curated,” suggesting a divide between data creators or users and curation practitioners in interpretation and understanding of how data should or could progress in the model.<sup>22</sup> Carlson’s emphasis on appraisal during the research process is, in many ways, unique to data curation. Could appraising data during the life cycle lead to a different outcome when compared to appraisal at the end of the life cycle? This is an area that the discipline of data curation should more fully explore.

Current trends in appraisal and selection methods in data curation have been built upon archival appraisal theories and collection development methods over the last decade.<sup>23</sup> The term *appraisal* refers to the method of identifying digital content’s permanent value for the purpose of long-term preservation. Therefore discussions of appraisal in data curation have been closely linked to institutional repository or data archival policies on collection development.<sup>24</sup> Appraisal criteria for initial selection decisions in a repository, for instance, function to maintain alignment with existing collections.<sup>25</sup> Early efforts to create data repositories were, largely, focused on a specific discipline or data type.<sup>26</sup> The rise of institutional data repositories and large-scale data publishing practices have expanded selection criteria and broadened existing collections beyond collection policies aimed

at a specific discipline, data type, or data format. Institutional data repositories, for example, collect, preserve, and give access to the research products of an entire institution, though they often arrange materials by department, college, or institute. Open, web-social repositories like figshare.com (<https://figshare.com>) and Dryad (<http://datadryad.org/>) continue to change the landscape of data repository options, allowing a greater variety of data to be accessible openly via the Web.

While similar to practices described in the library collection section, data curation focuses on digital contents rather than physical materials. Thus, a digital collection is measured by size, and its value can be based on the number of files, data sets, studies, and collections available in the repository. The usage metric for digital collections, namely the number of downloads, is still emerging as an assessment metric for the enduring value of data and is used in retention or disposition decisions.<sup>27</sup>

At the practical level, appraisal in data curation has developed to ensure that scientific records and data are usable over time. This is where, we believe, the two important issues (the value of data and the retention period) intersect and where it is important to address the question: What makes digital scientific records more or less usable? Although we lack standardized metrics to assess the value of data based on its reusability, there is a recent effort among data stewards to document and compile cases in which their openly shared research data is being reused by others.<sup>28</sup> This idea of reuse fuels the value assessment of data and drives the constantly evolving paradigm of federal-funder return on investment.

Another distinctive characteristic of data curation is the significant role that research communities play in appraising the value of data for long-term retention. In their data management plans, researchers may say that every data set should be preserved for the maximum period of retention (or forever, whichever comes first). We know, however, that due to resource concerns, the rapidly evolving technology environment, and changes in policy and authority, we cannot retain everything—sometimes the best we can hope for is planned obsolescence. From the researcher's perspective, appraisal criteria of scientific records and data should be biased toward relevance, significance, uniqueness, sensitivity, and the impact of their overall research output. These qualities are exactly those criteria at work in both library collections and records management. Communicating these perspectives, and the differences between them, should be a part of every retention and disposal discussion.

## Conclusion

Our review of these three disciplines—university records and information management, library collections, and data curation—suggests that there are criteria for data retention and destruction that go beyond a data set's projected value

over time. Additionally, we advise that anyone involved in deselection decisions also be aware of the local, legal, and disciplinary policies that impact data at each stage in the research life cycle. While data curation practices may enable data discovery and retrieval, maintain quality, add value, and facilitate reuse over time, perhaps curatorial “value-add” also incorporates the assessment of liability, risk, or resource cost over potential value. In these cases, the curation decision may lead to disposal of the data set. If the purpose of data curation is to add value at every stage of the research life cycle, we suggest that this definition includes the consideration of when to exit the life cycle. However, these decisions cannot be made at too high a level; like records management, the decision to dispose of a data set must take into account a variety of factors including (but not limited to) content, risk and liability, currency, scope, cost, quality, uniqueness, and external mandate. Not all of these factors will apply to every data set, but we believe that these criteria, combined with local practices, will provide a thorough basis for any decisions on when to exit the research life cycle.

## Notes

1. Although we were unable to identify any existing work that solely features the subject “death of data,” we have noticed that subscribers of *Research Data Management* discussion list, RESEARCH-DATAMAN hosted by Jisc, have actively (and in a timely manner for our book chapter) engaged in online discussions about related topics, including “data retention,” “identifying archival material,” and “retention of physical research data.” Threads on these topics are archived at: <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A1=ind1508&L=RESEARCH-DATAMAN#9> (threads in August 2015), <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A1=ind1509&L=RESEARCH-DATAMAN#12> (threads in September 2015), and <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A1=ind1510&L=RESEARCH-DATAMAN#33> (threads in October 2015).
2. Charlotte Brunskill and Sarah Demb, *Records Management for Museums and Galleries* (Oxford: Chandos Publishing, 2012); chapter 2 in this book provides the definition and practices of records management.
3. Jinfang Niu, “Appraisal and Selection for Digital Curation,” *International Journal of Digital Curation*, 9, no. 2 (2014): 68, doi:10.2218/ijdc.v9i2.272.
4. As suggested by Ross Harvey (“Appraisal and Selection,” in *DCC Digital Curation Manual*, ed. Seamus Ross and Michael Day [Edinburgh, UK: Digital Curation Centre, 2007], <http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/appraisal-and-selection>), ten appraisal criteria include value, physical condition, resources available, use, social significance, legal rights, format issues, technical issues, policies, and documentation. Niu actually develops her argument on four appraisal criteria: (1) mission alignment, (2) value of digital resources, (3) cost, and, (4) feasibility (Niu, “Appraisal and Selection,” 71–72).
5. Brunskill and Demb, *Records Management*; see chapter 7 for more detailed discussion on retention schedule and records management program.
6. OCLC has compiled a useful bibliography on weeding and deselection: OCLC, “Sus-

- tainable Collection Services: Weeding and Deselection Bibliography,” accessed May 29, 2016, <http://www.oclc.org/en-US/sustainable-collections/bibliography.html>.
7. Even from a more established preservation policy framework, ICPSR considers new digital content (e.g., website, audio, video, GIS) challenges, suggesting that existing policies, procedures, and practices need to be revised or re-engineered to encompass new digital content. See Inter-university Consortium for Political and Social Research (ICPSR), “ICPSR Digital Preservation Policy Framework,” April 2007, last revised June 2012, <https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/dpp-framework.html>.
  8. National Oceanic and Atmospheric Administration (NOAA), *NOAA Procedure for Scientific Records: Appraisal and Archive Approval, Guide for Data Managers*, (Washington, DC: NOAA, September 2008), [https://www.ngdc.noaa.gov/wiki/images/0/0b/NOAA\\_Procedure\\_document\\_final.pdf](https://www.ngdc.noaa.gov/wiki/images/0/0b/NOAA_Procedure_document_final.pdf).
  9. For an older but representative discussion of the lightly documented tension between basic administrative needs and scholarly concerns in university records management, see Don Skemer and Geoffrey Williams, “Managing the Records of Higher Education: The State of Records Management in American Colleges and Universities,” *American Archivist* 53, no. 4 (1990): 544–45, <http://www.jstor.org/stable/40293495>.
  10. Sharon Leslie and Ida Martinez, “Assessment and Weeding of a Clinical HIV/AIDS Collection in an Academic Library: A Case Study,” *Collection Management* 40, no. 3 (2015): 151, 153, doi:10.1080/01462679.2015.1040570.
  11. For more information on curation cost, see the project website of Collaboration to Clarify the Costs of Curation (4C) at <http://4cproject.eu/>. For more information on the cost analysis of digital collection, see the project called LIFE (Life Cycle Information for E-Literature), a collaboration between University College London (UCL) and the British Library, <http://www.life.ac.uk/>.
  12. On primary and secondary value for archives and records, see Theodore Schellenberg, “The Appraisal of Modern Public Records,” in *Modern Archives Reader: Basic Readings on Archival Theory and Practice*, ed. Maygene Daniels and Timothy Walch (Washington, DC: National Archives and Records Service, 1984), 68.
  13. William Saffady, *Records and Information Management* (Overland Park, KS: ARMA International, 2011), 64–65, provides a basic discussion of retention rationales.
  14. Rajia Tobia, “Comprehensive Weeding of an Academic Health Sciences Collection: The Briscoe Library Experience,” *Journal of the Medical Library Association* 90, no. 1 (2002): 94–98.
  15. Martha Kyrillidou, Shaneka Morris, and Gary Roebuck, “Rank Order Table 2: Titles Held,” in *ARL Statistics 2013–2014* (Washington, DC: Association of Research Libraries, 2014), 52.
  16. Mike Waugh, Michelle Donlin, and Stephanie Braunstein, “Next-Generation Collection Management: A Case Study of Quality Control and Weeding E-Books in an Academic Library,” *Collection Management* 40, no. 1 (2015): 19, doi:10.1080/01462679.2014.965864.
  17. *Ibid.*, 19–20.
  18. Tim Bucknall, Beth Bernhardt and Amanda Johnson, “Using Cost per Use to Assess Big Deals,” *Serials Review* 40, no. 3 (2014): 194–96, doi:10.1080/00987913.2014.949398.
  19. HathiTrust, “Access to Out-of-Print and Brittle or Missing Items,” accessed May 29, 2016, <https://www.hathitrust.org/out-of-print-brittle>.

20. Association of Southeastern Research Libraries, "Cooperative Journal Retention," accessed May 29, 2016, <http://www.aserl.org/programs/j-retain/>; Scholars Trust homepage, accessed June 17, 2016, <http://www.scholarstrust.org/>.
21. The most widely adopted definition of digital curation is provided by the Digital Curation Center, "What Is Digital Curation?" accessed June 17, 2016, <http://www.dcc.ac.uk/digital-curation/what-digital-curation>, and Inter-university Consortium for Political and Social Research (ICPSR), "Glossary of Social Science Terms," s.v. "Digital Curation," accessed June 17, 2016, <https://www.icpsr.umich.edu/icpsrweb/ICPSR/support/glossary#D>.
22. Jake Carlson, "The Use of Life Cycle Models in Developing and Supporting Data Services," in *Research Data Management: Practical Strategies for Information Professionals*, ed. Joyce M. Ray (West Lafayette, IN: Purdue University Press, 2014), 80.
23. Niu, "Appraisal and Selection," 66.
24. Angus Whyte and Andrew Wilson, "How to Appraise and Select Research Data for Curation," *DCC How-to Guides* (Edinburgh, UK: Digital Curation Centre, 2010), <http://www.dcc.ac.uk/resources/how-guides>.
25. *Ibid.*, 2.
26. Just as early efforts to create data repositories were, largely, focused on a specific discipline, different disciplines have required different approaches to appraisal and disposition. Esanu et al. (Julie Esanu, Joy Davidson, Seamus Ross, and William Anderson, "Selection, Appraisal, and Retention of Digital Scientific Data: Highlights of an ER-PANET/CODATA Workshop," *Data Science Journal* 3 (2006): 227–32, doi:10.2481/dsj.3.227dsj.3.227.) and Faundeen (John Faundeen, "Appraising U.S. Geological Survey Science Records," *Archival Issues: Journal of the Midwest Archives Conference* 32, no. 1 (2010): 7–22.) emphasize the importance of disciplinary-specific appraisal criteria.
27. There is a great body of work on download statistics focusing on institutional repositories, including Michael Organ, "Download Statistics: What Do They Tell Us?" *D-Lib Magazine* 12, no. 11 (2006), doi:10.1045/november2006-organ, and Stacy Konkiel and Dave Scherer, "New Opportunities for Repositories in the Age of Altmetrics," *ASIS&T Bulletin* 39, no. 4 (April/May 2013): 22–26, to name only a few.
28. In February 2016, two open data advocates from Innovations for Poverty Action and Mozilla Science Lab, Stephanie Wright and Stephanie Wykstra, have joined together to document examples of research data re-use from any scientific discipline: Stephani Wright, "Share Your Story of Research Data Re-use!" *Mozilla Science Lab* (blog), February 11, 2016, <https://www.mozillascience.org/share-your-story>.

## Bibliography

- Anson, Catherine, and Ruth R. Connell. *E-book Collections: SPEC Kit 313*. Washington, DC: Association of Research Libraries, 2009.
- Association of Southeastern Research Libraries. "Cooperative Journal Retention." Accessed May 29, 2016. <http://www.aserl.org/programs/j-retain/>.
- Blodgett, Peter, Jeremy Brett, Cathi Carmack, Anne Foster, Laura Uglean Jackson, Chela Scot Weber, Linda Whitaker, and Marcella Wiget. *Guidelines for Reappraisal and Deaccessioning*. Draft. Chicago: Society of American Archivists. July 12, 2011. <http://>

- www2.archivists.org/sites/all/files/GuidelinesForReappraisalAndDeaccessioning-DRAFT.pdf.
- Brunskill, Charlotte, and Sarah Demb. *Records Management for Museums and Galleries: An Introduction*. Oxford: Chandos Publishing, 2012.
- Bucknall, Tim, Beth Bernhardt, and Amanda Johnson, "Using Cost per Use to Assess Big Deals." *Serials Review* 40, no. 3 (2014): 194–96. doi:10.1080/00987913.2014.949398.
- Carlson, Jake. "The Use of Life Cycle Models in Developing and Supporting Data Services." In *Research Data Management: Practical Strategies for Information Professionals*, edited by Joyce M. Ray, 63–86. West Lafayette, IN: Purdue University Press, 2014.
- Digital Curation Center. "The Value of Digital Curation." Accessed May 29, 2016. <http://www.dcc.ac.uk/digital-curation/>.
- Duranti, Luciana. "Preface to the Special Issue on Data, Records, and Archives in the Cloud." *Canadian Journal of Information and Library Science* 39, no. 2 (2015): 91–96.
- Esanu, Julie, Joy Davidson, Seamus Ross, and William Anderson. "Selection, Appraisal, and Retention of Digital Scientific Data: Highlights of an ERANET/CODATA Workshop." *Data Science Journal* 3, 0 (2006): 227–32. doi:10.2481/dsj.3.227dsj.3.227.
- Faunden, John. "Appraising U.S. Geological Survey Science Records." *Archival Issues: Journal of the Midwest Archives Conference* 32, no. 1 (2010): 7–22.
- Harvey, Ross. "Appraisal and Selection." In *DCC Digital Curation Manual*. Edited by Seamus Ross and Michael Day. Edinburgh, UK: Digital Curation Centre, 2007. <http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/appraisal-and-selection>.
- HathiTrust. "Access to Out-of-Print and Brittle or Missing Items." Accessed May 29, 2016. <https://www.hathitrust.org/out-of-print-brittle>.
- Higgins, Sarah. "The DCC Curation Lifecycle Model." *International Journal of Digital Curation* 3, no. 1 (2008): 134–40. doi:10.2218/ijdc.v3i1.48.
- Inter-university Consortium for Political and Social Research (ICPSR). "ICPSR Digital Preservation Policy Framework." April 2007, last revised June 2012. <https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/dpp-framework.html>.
- Konkiel, Stacy, and Dave Scherer. "New Opportunities for Repositories in the Age of Altmetrics." *ASIS&T Bulletin* 39, no. 4 (April/May 2013): 22–26.
- Kyrillidou, Martha, Shaneka Morris, and Gary Roebuck. "Rank Order Table 2: Titles Held." In *ARL Statistics 2013–2014*, 52. Washington, DC: Association of Research Libraries, 2014.
- Leslie, Sharon, and Ida Martinez. "Assessment and Weeding of a Clinical HIV/AIDS Collection in an Academic Library: A Case Study." *Collection Management* 40, no. 3 (2015): 149–62. doi:10.1080/01462679.2015.1040570.
- Library of Virginia. "Records Management: Retention Schedules." Accessed May 29, 2016. <http://www.lva.virginia.gov/agencies/records/retention.asp>.
- National Archives. "Strategic Directions: Appraisal Policy." Accessed May 29, 2016. <http://www.archives.gov/records-mgmt/initiatives/appraisal.html>.
- National Oceanic and Atmospheric Administration (NOAA). *NOAA Procedure for Scientific Records: Appraisal and Archive Approval, Guide for Data Managers*. Washington, DC: NOAA, September 2008. [https://www.ngdc.noaa.gov/wiki/images/0/0b/NOAA\\_Procedure\\_document\\_final.pdf](https://www.ngdc.noaa.gov/wiki/images/0/0b/NOAA_Procedure_document_final.pdf).



- Niu, Jinfang. "Appraisal and Selection for Digital Curation." *International Journal of Digital Curation*. 9, no. 2 (2014): 65–82. doi:10.2218/ijdc.v9i2.272.
- OCLC. "Sustainable Collection Services: Weeding and Deselection Bibliography." Accessed May 29, 2016. <http://www.oclc.org/en-US/sustainable-collections/bibliography.html>.
- Office of Management and Budget. "Uniform Administrative Requirements for Grants and Agreements with Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations." OMB Circular A-110. November 6, 1999. [https://www.whitehouse.gov/omb/fedreg\\_a110-finalnotice](https://www.whitehouse.gov/omb/fedreg_a110-finalnotice).
- Organ, Michael. "Download Statistics: What Do They Tell Us?" *D-Lib Magazine* 12, no. 11 (2006). doi:10.1045/november2006-organ.
- Pennock, Maureen. "Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information." Article intended for publication in *Libraries and Archives* January 2007. Digital Curation Centre. 2007. [http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch\\_curation.pdf](http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf).
- Saffady, William. *Records and Information Management: Fundamentals of Professional Practice*. Overland Park, KS: ARMA International, 2011.
- Schellenberg, Theodore. "The Appraisal of Modern Public Records." In *Modern Archives Reader: Basic Readings on Archival Theory and Practice*. Edited by Maygene Daniels and Timothy Walch, 57–70. Washington, DC: National Archives and Records Service, 1984.
- ScholarsTrust. "Program Agreement: Association of Southeastern Research Libraries (ASERL) and Washington Research Library Consortium (WRLC) Shared Archives of Print Journal Collections." January 2013. [http://www.aserl.org/wp-content/uploads/2013/02/ASERL-WRLC\\_Shared\\_Archive\\_Print\\_Journals\\_AGREEMENT.pdf](http://www.aserl.org/wp-content/uploads/2013/02/ASERL-WRLC_Shared_Archive_Print_Journals_AGREEMENT.pdf).
- Skemer, Don, and Geoffrey Williams. "Managing the Records of Higher Education: The State of Records Management in American Colleges and Universities." *American Archivist* 53, no. 4 (1990): 532–47. Accessed May 29, 2016. <http://www.jstor.org/stable/40293495>.
- Society of American Archivists (SAA). "A Glossary of Archival and Records Terminology: Record." Accessed May 29, 2016. <http://www2.archivists.org/glossary/terms/r/record>.
- Tobia, Rajia. "Comprehensive Weeding of an Academic Health Sciences Collection: The Briscoe Library Experience." *Journal of the Medical Library Association* 90, no. 1 (2002): 94–98.
- Waugh, Mike, Michelle Donlin, and Stephanie Braunstein. "Next-Generation Collection Management: A Case Study of Quality Control and Weeding E-Books in an Academic Library." *Collection Management* 40, no. 1 (2015): 17–26. doi:10.1080/01462679.2014.965864.
- Whyte, Angus, and Andrew Wilson. "How to Appraise and Select Research Data for Curation." *DCC How-to Guides*. Edinburgh, UK: Digital Curation Centre. 2010. <http://www.dcc.ac.uk/resources/how-guides>.
- Wright, Stephanie. "Share Your Story of Research Data Re-use!" *Mozilla Science Lab* (blog), February 11, 2016. <https://www.mozillascience.org/share-your-story>.





## CHAPTER 11

# The Current State of Meta-Repositories for Data

*Cynthia R. Hudson Vitale*

## Introduction

Researchers have many options available to them in order to fulfill individual, funder, and publisher requirements to deposit and share research data. Thus many of their research outputs, including data, may be scattered across various institutional, domain, funder, publisher-supported, and general repositories and websites. Given this, a faculty member searching for data sets similar to his or her own research may find it difficult, if not impossible, to discover relevant sources. Without direct connections among the various research outputs, there are few mechanisms for anyone to understand what data, article, and code are related to the same research. This is a significant scholarly communications issue. Recently, much work has developed around online solutions to federate and link the records across these dispersed repositories, creating large meta-repositories of data.

Traditionally, in the scholarly literature meta-repositories of data have been categorized as digital libraries. What constitutes a digital library is complex, often defined ambiguously by the research community describing it.<sup>1</sup> When the World Wide Web was in its nascent stages, it too was considered a digital library. A more library-centric definition developed in the late 1990s, during which digital libraries were more closely tied to traditional libraries that had collection development plans, ensured the persistence of materials, preserved documents, and distributed the resources.<sup>2</sup> While meta-repositories of data fit this definition, they also have

a number of distinct qualities that set them apart, including a close focus on research materials and the aggregation of metadata or data from dispersed sources.

A previous study of digital libraries that are more similar to these meta-repositories of data, compiled by the Digital Repository of Ireland (DRI), focused on how digital objects were being cared for internationally.<sup>3</sup> The authors found three different models: the metadata aggregator, the single-site repository, and the multi-site repository. DRI also indicated that funding agencies place a greater emphasis on access rather than preservation of the digital content, which may ultimately put the ongoing availability of content at risk.

Extending the work completed by DRI, this chapter comparatively analyzes the major international meta-repositories of data to better understand their goals and missions, overlaps in services and content, and any common challenges.

## Community Initiatives and Solutions to Support Meta-Repositories of Data

Though the scholarly literature around meta-repositories of data is not extensive, a number of international organizations have become more inclusive of data repository agendas by establishing working groups to address repository technical issues, metadata challenges, and interoperability.

Founded in 2009, the Confederation of Open Access Repositories (COAR) seeks to create community and support for repositories worldwide. Current members include the Vienna University Library and Archive Services, the University of Antwerp, McMaster University Library, bepress, and the World Bank, to name a few.<sup>4</sup> The COAR organization and community builds capacity, aligns policies and practices, and acts as a global voice for the repository community. COAR's approximately 100 members represent libraries, universities, research institutions, government funders, and others. According to COAR's 2016–2018 strategic plan, one of its primary objectives is to work towards interoperability with research data management repositories and systems.<sup>5</sup> Interoperability work such as this might allow federated data repositories to more easily aggregate metadata records and exchange information.

The Research Data Alliance (RDA) was established in 2013 as a grass-roots organization that builds the technical and socio-technical infrastructure for data sharing.<sup>6</sup> It is organizationally comprised of approximately sixty-two interest and working groups that include focus on everything from wheat interoperability to sharing sensitive data and developing a data type registry, to name a few. One community group that includes repository interoperability among its goals is the

Repository Platforms for Research Data Interest Group.<sup>7</sup> A deliverable of this group is to create a matrix of functional requirements related to repository platforms, which may also relate to specifications for a generic application programming interface. A newly proposed group, titled Research Data Repository Interoperability, is looking specifically at research data repository interoperability as a working group. The main objectives of this group are to identify, evaluate, and establish standards for interoperability between different research data platforms. Already, repository developers representing DSpace, Hydra, Fedora, DataOne's Metacat, and others have agreed to implement these recommendations upon the close of the working group.<sup>8</sup> These types of community-developed and -initiated projects ensure wide adoption and solutions that fit the needs.

Organizations that support the quality and accessibility of data are not new. The International Council for Science: Committee on Data for Science and Technology (CODATA) is an organization established over forty years ago. One of CODATA's main objectives is to facilitate international cooperation among those institutions collecting, organizing, and using data.<sup>9</sup> This work is primarily facilitated through the committees and working groups focused on projects of specific scope, such as legal interoperability.

Finally, the International Council for Science: World Data System (ICUS/WDS) is a unique organization that promotes universal access and long-term stewardship of quality-assured scientific data and data services products.<sup>10</sup> This organization, comprised of working groups, is unique because it also provides services and aggregates data from member organizations, thereby acting as a "meta-repository."

Meta-repositories of data participate in, support, and are putting into practice many of the recommendations and outputs developed or in development by these community initiatives. Yet understanding how these meta-repositories of data work together, overlap, or complement each other has not been examined. Thus, the goal of this study is to comparatively analyze these systems in order to better understand the current state of meta-repositories for data.

## Methods

A unified term to describe meta-repositories of data currently does not exist, which makes conducting Web searches to identify these systems impossible. Conducting Web searches using the terms *federated repositories* and *repository aggregator* resulted in zero relevant systems. Thus, the meta-repositories of data described here were primarily identified through the author's knowledge of such systems and suggestions from colleagues.

Thirteen meta-repositories were chosen for analysis based upon the following criteria:

1. Content: The meta-repositories of data were receiving or harvesting data (either metadata or digital data objects) from individual repository platforms.
2. Language: The meta-repositories of data websites were written in English.
3. Spatial: International repository aggregators were within scope of this analysis.

The thirteen repositories are listed in table 11.1.

**TABLE 11.1**  
The Meta-Repositories Chosen for Analysis in this Study

Meta-Repository	Mission	URL
1. Australian Research Data Commons (ANDS)	ANDS is a system built and maintained in Australia to <ul style="list-style-type: none"> <li>• “make Australian research data collections more valuable by managing, connecting, enabling discovery and supporting the reuse of this data”</li> <li>• “enable richer research, more accountable research; more efficient use of research data; and improved provision of data to support policy development.”<sup>a</sup></li> </ul>	<a href="http://ands.org.au">http://ands.org.au</a>
2. Beilefeld Academic Search Engine (BASE)	BASE is a portal established by Bielefeld University Library, United Kingdom that integrates Open Archives Initiative (OAI) resources as one information type among others into the local digital library environment, together with catalogs, article databases, and digitized collections.	<a href="https://www.base-search.net/">https://www.base-search.net/</a>
3. Connecting REpositories (CORE)	CORE is a UK-based meta-repository that seeks “to aggregate all open access research outputs from repositories and journals worldwide and make them available to the public.” <sup>b</sup>	<a href="https://core.ac.uk/">https://core.ac.uk/</a>
4. Data.gov	Data.gov is the home of US government metadata. Non-federal data sources can also be added to the data set voluntarily.	<a href="http://www.data.gov/">http://www.data.gov/</a>

**TABLE 11.1** (continued)

Meta-Repository	Mission	URL
5. Data Archiving and Networked Services (DANS)	Developed in the Netherlands, DANS is a service institute that promotes sustained access to digital research data.	<a href="http://www.dans.knaw.nl/en">http://www.dans.knaw.nl/en</a>
6. DataBridge	DataBridge is a cross-institutional collaboration that aims to make the “long tail” of data more discoverable.	<a href="http://databridge.web.unc.edu/">http://databridge.web.unc.edu/</a>
7. DataCite	DataCite is an organization that works with data centers to assign digital object identifiers to research assets.	<a href="https://www.datacite.org">https://www.datacite.org</a>
8. EUDAT	EUDAT is a system that includes data access, deposit, sharing, archiving, identification, and discovery of research data produced across the European Union.	<a href="https://eudat.eu">https://eudat.eu</a>
9. ICSU/World Data System (WDS)	Launched in Japan, ICSU/WDS research data system seeks to enable universal and equitable access to scientific data.	<a href="https://www.icsu-wds.org">https://www.icsu-wds.org</a>
10. OpenAIRE	Initiated in the European Union, OpenAIRE brings together scholarly metadata to support open scholarship and improve the reuse of publications and data.	<a href="https://www.openaire.eu/">https://www.openaire.eu/</a>
11. OpenDOAR	OpenDOAR is a directory of open-access academic repositories.	<a href="http://opendoar.org/">http://opendoar.org/</a>
12. OneRepo	OneRepo is a system that seeks to bring together all open-access scholarly articles.	<a href="http://onerepo.net">http://onerepo.net</a>
13. SHARE	SHARE is a metadata data set about research and scholarly activities through the research life cycle (such as data management plans, funder information, articles, data sets, etc.)	<a href="http://share-research.org">http://share-research.org</a>

a. “About Us,” Australian National Data Service, accessed May 26, 2016, <http://www.ands.org.au/about-us>.

b. “About CORE,” CORE homepage, accessed May 26, 2016, <https://core.ac.uk/>.

It should be noted, that while LaReferencia is a known meta-repository for South America, the website is entirely in Spanish. Although OpenDOAR is a directory of open-access repositories, it also includes a Google search widget that allows a user to search across the content of the repositories it indexes; thus, it was included in this study.

The comparative analysis was conducted by evaluating the websites of each meta-repository of data across fifteen variables in four distinct areas (see table 11.2) with the goal of better understanding the repository aggregator's background, content coverage, metadata employed, and functionality of the search interface. All data for the analysis was manually collected during the period October 10, 2015–April 7, 2016. The author searched primarily through each website's About pages and search interfaces and used white papers and other website documents to collect the comparative data. The raw data, along with hyperlinks to the document where the information was collected from, is available in the data set that accompanies this chapter.

**TABLE 11.2**  
**The Meta-Repository Website Analysis Used Variables**  
**Categorized into Four Areas**

Area	Variables Collected
Background	Date founded, goals/vision, mission, funding model
Content Coverage	Time span, spatial/geographic parameters, domain specificity, data types, providers, number of records, update frequency
Metadata	Standards, elements
Functionality	Faceted searching, feeds/alerts

## Results

The results of the website analysis show various points of similarities among the thirteen meta-repositories of data. Six of the meta-repositories were created to support national missions to ensure quality data and accessibility (meta-repositories 1, 4, 5, 8, 9, 10), while the remaining were created as responses to growing scholarly communication needs, to maximize research impact, and to otherwise promote science. For example, the mission of the ANDS is to make Australian research data collections more available “by managing, connecting, enabling discovery and supporting the reuse of this data.” In contrast, SHARE’s mission is “to maximize research impact by making a comprehensive inventory research widely accessible, discoverable, and reusable.”

All of the repository aggregators analyzed, except for BASE (established in 2004), were established within the last ten years, with the majority ( $n=8$ ) established or founded within the last six years (meta-repositories 3, 4, 6, 7, 8, 10, 12, 13). The repository aggregators fell into four distinct funding categories: those that are federally or nationally funded ( $n=6$ ), commercially and organizationally

funded ( $n=4$ ), grant funded ( $n=2$ ), and one that is currently seeking funding or whose funding is unsecured ( $n=1$ ).

## Content

From a content perspective, the majority of the meta-repositories were harvesting content from repositories worldwide ( $n=9$ ), while two were limited to nations and one was unknown in spatial coverage. None of the repository aggregators were limited to a specific domain (i.e., gathering source information only from a specific scientific discipline). While all of the repository aggregators had metadata about data sets in their systems, many also had articles, theses and dissertations, and conference papers and presentations. One repository aggregator, OpenDOAR, also included content such as audiovisual material and learning objects. Most systems were simply aggregating the metadata, but a handful of the meta-repositories had the actual digital asset stored, including CORE and Data.gov.

The number of providers varied significantly among the meta-repositories, ranging from 20 (OneRepo) on the low end to over 6,000 on the upper end (CORE). There was a low, but surprising, amount of overlap found among the institutional repositories covered within these systems. Of the thirteen repository aggregators, only five made their provider list available. Of these, over 1,400 repositories were aggregated overall. While no deduplication was completed as part of this analysis, these aggregators have collectively brought together millions of records. Individually, some of the aggregators did not release how many records they had (OpenDOAR and OneRepo). Time spans of the content found in the aggregated repositories were also difficult to determine. BASE had the longest known temporal span, with records available for materials created in the 1000s.

## Functionality

In regard to search features and faceting found in the meta-repositories for data, all working systems had some type of advanced search limiters. The most common types of features were facets that allowed the user to limit the results by a subject area, institution, or publication year. The Australian National Data Service had a unique function that allowed the user to find related people and related organizations from a search query.

Conducting a search, having appropriate results, and accessing the data set are a primary goal of these systems, but being able to download the metadata of the search results or export metadata in some manner was investigated as well. Seven of the meta-repositories for data had a tool to allow the user to export search results or access the underlying metadata for records in the repository. These tools ranged in implementation from e-mail alerts and SPARQL endpoints

to more robust APIs. Two of the systems, SHARE and OpenAIRE, were developing alerting tools for searches. These tools, such as, SHARE Notify,<sup>11</sup> allowed the user to conduct a search across the SHARE data set and set up an atom feed to receive real-time updates. Use cases for this tool are many, but include the ability to stream this data to a web interface that would keep researchers up-to-date on relevant scholarship or alert local institutional repositories about new faculty-created materials available for harvesting. The Literature Broker Service, in development at OpenAIRE, is similar to the latter Notify use case. It is a subscription-based system that aims to support institutional repository managers by altering them to new publication objects not currently in their collections. This system has the added benefit of disseminating additional or updated metadata related to records already in the repository.<sup>12</sup>

## Metadata

One of the most glaring areas where many of the meta-repositories for data systems did not align was in their use of metadata standards. Of the thirteen systems, only two used the same standard: DataCite and OpenAIRE (DataCite metadata standard). The remaining eleven systems all used a local standard—RIOXX, DDI Lite, panFMP, DDI, RIF-CS—or were not using a standard for various reasons. At one end of the spectrum was one system, EUDAT, that required only one metadata element for creating a record: a title. On the opposite end, DataBridge required the most metadata with over twenty-four elements from the DDI Lite standard. The most common elements found in meta-repository metadata schemes were title ( $n=10$ ) and author/contributor ( $n=6$ ).

## Discussion

This comparison revealed varying stages of development for each meta-repository. Many were just recently launched in the last five years, which means their systems may not have undergone many iterations to improve functionality or usability. Additionally, as many of these repositories overlap in content and mission, the ongoing availability is of concern. Federal and grant funds are often limited, thus many of these systems may be competing for the same funding streams.

The metadata issue is also incredibly significant. Without a common standard and element set it is doubtful that these systems will be fully interoperable. This issue is not limited to just meta-repositories; Moulaison, Dykas, and Galant found that roughly half of the twenty-three open-access repositories they surveyed were using the same metadata standard, Dublin Core.<sup>13</sup> The remaining used a combination of Qualified Dublin Core, MODS, and MARC. Additional-



ly, given the flexibility of many of these standards, the application of a standard varies both within and across systems. For example, dates can represent vastly different points given how a local repository makes use of the field. A date field can be interpreted as the date the asset was published online, the date the asset was created, and the date it was published in print. When a meta-repository pulls these two repositories together in the same system, the inconsistency is problematic. The answer to these issues might be to use computer systems to parse and normalize, or for the data repository community to come together and agree on a more rigid application of the metadata elements in a standard.

Additionally, while many of the repository aggregator missions were to support the accessibility and persistence of scholarship, few of the repository aggregators had facets that allowed users to limit to open-access materials. Although they claimed persistence as a priority, how this was facilitated was not evident across any of the meta-repositories for data. For example, none of the meta-repositories assigned DOIs or persistent identifiers to the metadata records they were aggregating, and few, if any, had curation treatment procedures in place for the metadata. Policies of how the meta-repository handles withdrawn records are not always evident.

Finally, many of these meta-repositories of data have come to act as *de facto* representatives of the smaller systems they aggregate or harvest from. Much like a traditional consortium, the meta-repositories can advocate for the interests of the other systems, recommend metadata standards, suggest best practices for metadata element values, and potentially create inventories of technical infrastructure for data repositories.

## Conclusion

Scholarly communication is in need of systems to pull together and link dispersed research objects. Just as Netflix revolutionized film discovery and rental, meta-repositories are needed to discover and highlight research from varying providers, make recommendations, show relationships between research and researchers, and make connections among the digital assets. The whole story of research, and the complete scholarly record, is more than just the final publication. It includes funder information, data sets, documentation, and code in many cases. The meta-repositories of data are one tool that seeks to address this issue. There exist many challenges to making these systems robust and operational enough to fit the scholarly communications need. Community involvement at the local level is integral to ensuring the success of these systems. Engagement with COAR, RDA, CODATA, or even the meta-repositories directly, ensures the ongoing viability of these useful systems.

## Notes

1. Peter J. Nurnberg, Richard Furuta, John J. Leggett, Catherine C. Marshall, and Frank M. Shipman III, "Digital Libraries: Issues and Architectures," in *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries. Austin, Texas, June 11-13, 1995* (Austin, TX: 1995), 147–53, <http://www.jcdl.org/archived-conf-sites/dl95/papers/nuernberg/nuernberg.html>.
2. Donald J. Waters, "What Are Digital Libraries?" *CLIR Issues*, no. 4 (July/August 1998), <http://www.clir.org/pubs/issues/issues04.html>.
3. A. O'Carroll, S. Collins, D. Gallagher, J. Tang, and S. Webb, *Caring for Digital Content: Mapping International Approaches* (Maynooth: NUI Maynooth; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy, 2013), <http://dri.ie/caring-for-digital-content-2013.pdf>.
4. Confederation of Open Access Repositories, "About COAR: Strategic Plan," accessed November 11, 2015, <https://www.coar-repositories.org/about/coar-ev/strategic-plan/>.
5. Confederation of Open Access Repositories, *COAR Strategy 2016–2018 and COAR Work Plan 2016–2017*, final version (Göttingen, Germany: COAR, November 5, 2015), <https://www.coar-repositories.org/files/COAR-Strategy-2016-2018-Final.pdf>.
6. Research Data Alliance homepage, accessed November 15, 2015, <http://rd-alliance.org>.
7. Research Data Alliance, "Repository Platforms for Research Data: Case Statement," accessed November 20, 2015, <https://rd-alliance.org/group/repository-platforms-research-data/case-statement/repository-platforms-research-data-case>.
8. Research Data Alliance, "Research Data Interoperability Case Statement," Google Doc, Accessed March 29, 2016, [https://docs.google.com/document/d/11XZBXIxSOE\\_d0n-1JsaYx2LhRwkwel5OQNHBaXKG-a\\_o/edit](https://docs.google.com/document/d/11XZBXIxSOE_d0n-1JsaYx2LhRwkwel5OQNHBaXKG-a_o/edit).
9. Committee on Data for Science and Technology (CODATA), "Our Mission," accessed November 1, 2015, <http://www.codata.org/about-codata/our-mission>.
10. World Data System, "About," accessed November 1, 2015, <https://www.icsu-wds.org/organization>.
11. Tyler Walters and Judy Ruttenberg, "SHared Access Research Ecosystem," *Educause Review* 49, no. 2 (2014), 56–57.
12. Michele Artini, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi, and Andrea Mannocci, "The OpenAIRE Literature Broker Service for Institutional Repositories," *D-Lib Magazine* 21, no. 11/12 (November/December 2015), doi:10.1045/november2015-artini.
13. Heather Lea Moulaison, Felicity Dykas, and Kristen Gallant, "OpenDOAR Repositories and Metadata Practices," *D-Lib Magazine* 21, no. 3–4 (March/April 2015), doi:10.1045/march2015-moulaison.

## Bibliography

- Artini, Michele, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi, and Andrea Mannocci. "The OpenAIRE Literature Broker Service for Institutional Repositories." *D-Lib Magazine* 21, no. 11/12 (November/December 2015). doi:10.1045/november2015-artini.

- Committee on Data for Science and Technology (CODATA). "Our Mission." Accessed November 1, 2015. <http://www.codata.org/about-codata/our-mission>.
- Confederation of Open Access Repositories. "About COAR: Strategic Plan." Accessed November 15, 2015. <https://www.coar-repositories.org/about/coar-ev/strategic-plan/>.
- . *COAR Strategy 2016–2018 and COAR Work Plan 2016–2017*. Final version. Göttingen, Germany: COAR, November 5, 2015. <https://www.coar-repositories.org/files/COAR-Strategy-2016-2018-Final.pdf>.
- Moulaison, Heather Lea, Felicity Dykas, and Kristen Gallant. "OpenDOAR Repositories and Metadata Practices." *D-Lib Magazine* 21, no. 3–4 (March/April 2015). doi:10.1045/march2015-moulaison.
- Nurnberg, Peter J., Richard Furuta, John J. Leggett, Catherine C. Marshall, and Frank M. Shipman III. "Digital Libraries: Issues and Architectures." In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries. Austin, Texas, June 11–13, 1995*, 147–53. Austin, Texas, 1995, <http://www.jcdl.org/archived-conf-sites/dl95/papers/nuernberg/nuernberg.html>.
- O'Carroll, A., S. Collins, D. Gallagher, J. Tang, and S. Webb. *Caring for Digital Content: Mapping International Approaches*, Maynooth: NUI Maynooth; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy, 2013. <http://dri.ie/caring-for-digital-content-2013.pdf>.
- Research Data Alliance. "Repository Platforms for Research Data: Case Statement." Accessed November 20, 2015. <https://rd-alliance.org/group/repository-platforms-research-data/case-statement/repository-platforms-research-data-case>.
- . "Research Data Interoperability Case Statement." Google Doc. Accessed March 29, 2016. [https://docs.google.com/document/d/11XZBXIxSOE\\_d0n1JsaYx2LhRwkwel5OQNHBaXKG-a\\_o/edit](https://docs.google.com/document/d/11XZBXIxSOE_d0n1JsaYx2LhRwkwel5OQNHBaXKG-a_o/edit).
- Research Data Alliance homepage. Accessed November 15, 2015. <http://rd-alliance.org>.
- Walters, Tyler, and Judy Ruttenberg. "SHared Access Research Ecosystem." *Educause Review* 49, no. 2 (2014): 56–57.
- Waters, Donald J. "What Are Digital Libraries?" *CLIR Issues*, no. 4 (July/August 1998). <http://www.clir.org/pubs/issues/issues04.html>.
- World Data System. "About." Accessed November 1, 2015. <https://www.icsu-wds.org/organization>.





## CHAPTER 12\*

# Curation of Scientific Data at Risk of Loss

## Data Rescue and Dissemination

*Robert R. Downs and Robert S. Chen*

Data rescue offers an opportunity for digital repositories, including institutional repositories, data archives, and scientific data centers, to provide access to potentially valuable scientific data that is at risk of being lost. Rescue may be valuable not only to restore access to data of past scientific interest, such as environmental observations or social surveys, but also to recover historic information about the state of knowledge and science at the time the data was collected or assembled. Scientific data may need to be rescued at any stage along the data life cycle, and the extent of data curation that was completed prior to a data rescue effort may vary, depending on the circumstances that led to the need for data rescue. The level of effort required to complete a data rescue depends largely on the condition of the data being rescued, the availability and quality of data documentation and provenance information, and the accessibility of the data producers. In extreme cases, data organization and documentation are poor, and those knowledgeable about how the data was collected or developed are no longer available. In some cases, collections of data sets may need to be rescued from an existing archive that is no longer sustainable. In short, scientific data may be at risk of loss for a variety of reasons, and a data rescue effort can present new challenges for data curation and dissemination operations.

---

\* Copyright The Trustees of Columbia University in the City of New York, 2016. Creative Commons Attribution 4.0 License, CC BY (<https://creativecommons.org/licenses/by/4.0/>).

We report here on a recent effort by the NASA Socioeconomic Data and Applications Center (SEDAC) to rescue the Millennium Ecosystem Assessment (MA) collection of scientific data as a case study on the issues raised by a data rescue effort from an existing archive that had not fully curated the original data. The MA was an international survey of the world's ecosystems conducted by the scientific community in 2001–2005 involving more than 1,300 experts from around the world. As part of the MA, a diverse set of environmental and socioeconomic data was assembled and integrated in order to enable scientific analysis and assessment in support of policy and decision making. This data was held by the US Geological Survey (USGS) National Biological Information Infrastructure (NBII), which was terminated by the US government in early 2012.<sup>1</sup> This case study describes what happened to the data after the MA was completed, why data rescue was subsequently needed, the process used to decide on the data rescue effort, and the subsequent issues and challenges addressed in rescuing the MA data. The core preservation need for the MA collection is described along with the tradeoffs involved in conducting the data rescue. Based on the case study, we summarize lessons learned from the data rescue effort, including lessons for projects that create or collect data, for repositories that acquire data from such projects, and for those engaged in rescuing data. Of course, whether there will be significant scientific or historical benefit resulting from this rescue effort remains to be seen.

## Benefits of Data Rescue

Data repositories that work closely with the scientific community are likely to encounter opportunities to conduct data rescue activities that could contribute to science by facilitating the use of legacy data for new studies. The term *data rescue* refers to efforts that enable the sustained use of data that otherwise might go unused. The World Meteorological Organization has defined data rescue as “the ongoing process of 1. preserving all data at risk of being lost due to deterioration of the medium and; 2. digitizing current and past data into computer compatible form for easy access.”<sup>2</sup>

Data rescue needs to occur before the data in question becomes completely inaccessible or unusable, and ideally should occur while those scientists or others familiar with the data are still available to provide important information about the data, its origin, collection, and management, and its quality. Data rescue can enable studies that would not otherwise be possible without the rescued data.<sup>3</sup> For example, legacy data can fill gaps about events and anomalies that might not be part of a longitudinal study. In summarizing several data rescue efforts, Griffin noted that “legacy data may be the best, sometimes the only, sources of information about those critical departures from the norm.”<sup>4</sup> As another example, data rescued from various publications of 1855 and 1856 and from weather station

records of the era, along with other sources, has revealed extreme precipitation events that occurred during that period in the Iberian Peninsula.<sup>5</sup>

Scientific data rescue efforts also offer opportunities for repositories to improve their collections and contribute to the infrastructure, advancement, and application of science. Climate records for countries in the Mediterranean region from the past few centuries are currently being inventoried and rescued to facilitate longitudinal climate assessments and predictions.<sup>6</sup> Many important long-term climate data series have been developed from historical records, such as those available from the Climatic Research Unit at the University of East Anglia, the Climate Data Library of the International Research Institute for Climate and Society (IRI), and the US National Climatic Data Center (NCDC).<sup>7</sup> This data has been critical not only to the advancement of science, but also to international assessments conducted by the Intergovernmental Panel on Climate Change (IPCC).<sup>8</sup>

Rescue also may be used to recover historic information about the state of knowledge and science at the time the data was collected or assembled. For example, historians or political scientists may be interested in understanding the level of scientific awareness and understanding at important points in decision making that requires significant scientific input.<sup>9</sup> Another possible benefit results when the cost of the data rescue represents a fraction of the cost of any new data collection.<sup>10</sup> In such cases, data rescue could offer an efficient alternative to new data acquisition, saving time and money.

## Challenges of Data Rescue for Repositories

A data rescue effort offers unusual challenges for repositories, such as scientific data centers and archives, which routinely work with data producers and user communities to curate data and improve its potential for use by the communities that they serve. A data rescue could be required as a result of various circumstances, such as media decay and obsolescence, laboratory closure, absence of documentation and data quality information, non-digital data capture, and missed opportunities to capture data within a data management system.<sup>11</sup> Data rescue efforts can be quite diverse, reflecting the different kinds of data that have been collected, the effects of time and technological change, and the availability of resources for obtaining the data and enabling its sustained use by an identified community. Complex data rescue efforts can involve developing automatic correction and conversion methods for recovering data, for example from multiple satellite instruments or creating metadata from forty-year-old tapes to study sea ice during the 1960s.<sup>12</sup> Furthermore, data rescue could require collection, digitization, and quality control of historical data from various sources that are

no longer publicly available, including historical records from obsolete analog instruments and handwritten observations obtained from historical documents, such as ship logbooks and signal stations that create a comprehensive time series climate record.<sup>13</sup> Most of these situations mean that normal processes for properly managing the life cycle of scientific data cannot be carried out in a routine manner due to inadequate data management during parts of the data life cycle.

Often, knowledge about the context of the data being rescued is not readily available. Ideally, such knowledge can be gathered from publications or technical documents describing the data, or else obtained from members of the original study team or others intimately familiar with the data. For example, handwritten materials that have faded or are illegible pose challenges that can be mitigated if members of the original data collection team can help interpret the materials or fill in the information gaps.<sup>14</sup> Furthermore, Knapp, Bates, and Barkstrom warned “that without the active participation from the complete chain of data provider, archive, and users, data sets will atrophy and become unusable.”<sup>15</sup> However, when a decision has been made to rescue a particular data set or collection, the rescuing repository may not know about relevant sources of information and may not be aware of who was involved in creating and managing the data or how to reach them—assuming they are still available to be reached!

In the absence of complete information about the data in need of rescue or assistance from those who possess knowledge of the data and its provenance, a data rescue effort may require divergence from rigorous data curation and quality assurance practices, such as those that are usually completed within a scientific data center. In cases where information about scientific data and its quality is limited, tradeoffs may be necessary to balance the desire for scientific rigor or completeness, the requirements of potential uses and users, and the available resources at hand. The adoption and use of specialized hardware and software may be needed, and the required capabilities for conducting a data rescue could be different for each data set in need of rescue. Furthermore, data rescue efforts in developing countries, even though they could be of significant value, are prone to conditions that pose risks for data preservation (even for current data management efforts), and developing countries typically do not have the resources to conduct data rescue efforts.<sup>16</sup>

## Repository Considerations for Data Rescue

Scientific data may need to be rescued at any stage along the data life cycle, and the extent of data curation that was completed prior to the data rescue effort may vary. Whereas some data rescue initiatives involve digitization of data from analog form, rescue of data from the last half-century can involve remastering to



convert digital data from older databases, formats, and media.<sup>17</sup> The condition of the data and associated documentation that are in need of rescue will likely affect the level of effort required to make the rescued data usable. For example, significant effort may be necessary when data values have not been properly collected into a data set or curated, and the data producers are no longer available. On the other hand, a properly curated and usable collection of data rescued from an archive that is no longer sustainable may take only a small amount of effort to ingest and assimilate into a new repository.

Although it might be ideal to bring older or orphaned data sets up to current standards of data management, doing so could consume resources that are needed to manage current data that could have many more users, uses, and scientific or societal benefits. In this case, consider a basic data rescue strategy that includes digital preservation of the data files, identification and preservation of *critical* documentation, and preparation of appropriate preservation and discovery metadata. While development of *complete* documentation would be ideal, a high priority for documentation should be the identification of data ownership information and, if possible, securing of dissemination rights from the owners if the owners can be identified and reached. This strategy ensures that data is not lost forever; on the other hand it leaves some onus on future users to invest time and effort to obtain any additional information about the data needed to interpret and use the data appropriately to meet their own objectives.

Observations from a data rescue effort by a scientific data center can help inform future data rescue efforts in their decision-making process. This case study of a data rescue effort, which was completed in 2015 by the NASA Socioeconomic Data and Applications Center (SEDAC),<sup>18</sup> provides insight into the issues, challenges, and choices that future data rescue efforts might encounter. SEDAC routinely acquires, manages, preserves, and prepares data about human interactions in the environment for dissemination to scientific communities, decision makers, and the public. The case study describes how the collection of data was identified, assessed, and selected for the data rescue effort. The workflow of the data rescue, including planning, preparation, organization, review, and dissemination of the collection, is also described. Successful aspects of the described data rescue are discussed to inform future data rescue efforts and to suggest opportunities for repositories to plan for and complete their own data rescue efforts.

## Rescue of the Millennium Ecosystem Assessment (MA) Data

The Millennium Ecosystem Assessment (MA) data was developed as part of a worldwide appraisal of ecosystems and conducted under the auspices of the Unit-

ed Nations by more than 1,300 scientists between 2001 and 2005. The data was gathered from multiple sources and assembled for analysis, forming the basis for a series of influential reports on the state of the world's ecosystems issued in 2005.<sup>19</sup> The data included version 3 of SEDAC's Gridded Population of the World (GPW) data set as well as "alpha" versions of several other SEDAC data sets that were made available to the MA in advance of formal release. All of the data was originally held by the National Biological Information Infrastructure (NBII) program of the United States Geological Survey (USGS). However, the US Congress cut the budget for NBII beginning in the 2012 federal fiscal year, leading to closure of the NBII's main website and associated nodes in January 2012.<sup>20</sup>

At that time, SEDAC recognized that there was scientific and historical value in the MA collection of data, and that this data was at high risk of being permanently lost due to the NBII's termination. Several SEDAC staff members had been involved in the MA and the NBII, and were therefore knowledgeable about the origins of the data and who had been involved. An initial assessment was conducted to determine the relevance of the data to specific SEDAC mission objectives and to meeting future user needs. It was determined that the socioeconomic scenarios developed for the MA would be of high interest to SEDAC users and that other MA data could be of interest to user communities concerned with climate impacts, adaptation, and vulnerability; environmental sustainability; agricultural and forest productivity; and land use and land cover change.

SEDAC acquired copies of the MA data in 2012 from an individual who had worked with the NBII for a preliminary review. The initial inventory of the collection identified 43 possible data products in 92 data files, with a total volume of approximately 1.75 gigabytes. The files were not well-documented and did not include any data set-level metadata or permissions documentation, reflecting the limited attention given to formal data management during the MA. Additional documentation, provenance information, and methodological details for the data were sought by e-mail from members of the data creation teams, with limited success. Many MA scientists were not available or had limited recollection of specific information about the collection contents. SEDAC determined that it would take substantial staff time (over multiple years) to archive and document all of the 43 data products individually with appropriate provenance and context information, and that in some cases important information might not be recoverable. In most cases, data had been superseded by more recent versions, so the primary interest in the data would be historical.

In light of these factors, and considering its other data development, management, and dissemination priorities, SEDAC decided to propose a basic data rescue effort that could enable future discovery and use of the MA collection. In May 2013, the SEDAC User Working Group (UWG), an advisory group of scientists, representative users, and other experts that meets annually,<sup>21</sup> approved

SEDAC's plan to archive and disseminate the MA collection with limited additional value-added efforts.

To streamline the data rescue effort, SEDAC organized the MA files thematically into six data sets for online dissemination: MA Biodiversity, MA Climate and Land Cover, MA Ecosystems, MA Population, MA Rapid Land Cover Change, and MA Scenarios. These six MA data sets contain the original MA files in their original formats with supplementary information obtained from various sources. SEDAC staff members worked intensively to clarify authorship and dissemination rights, working with the relevant report or chapter authors. However, SEDAC decided to refer users to the published MA assessment reports for detailed information on the scientific background of the data and its use in the MA analysis. The data and the MA assessment reports were analyzed to create a collection description and a summary and metadata record for each of the MA data sets.

Prior to dissemination, each data set in the MA collection was accessed and analyzed to ensure that the data quality was not compromised and that the data could be accessed by interested users. Each data set also received an internal "alpha" review by SEDAC scientists and staff, followed by a "beta" scientific and technical review by selected external users including members of the SEDAC UWG. The SEDAC Configuration Management Board (CMB) reviewed all comments received and ensured that corrections to collection and data set descriptions and to metadata were completed prior to public release. Each data set in the MA collection was archived to ensure preservation prior to dissemination.

## Dissemination of the Millennium Ecosystem Assessment (MA) Data

Within the structure of the SEDAC website, a data collection was established to provide access to the MA collection (<http://sedac.ciesin.columbia.edu/data/collection/ma>). The collection description on the MA collection webpage explains that as the result of "a data rescue effort, minimal documentation and support is provided,"<sup>22</sup> to notify potential users that the data sets in the MA collection might not meet their expectations. As for other SEDAC collections, the MA collection webpage then links to the landing page for each data set (see figure 12.1), which contains a data set description and a recommended data citation, including an assigned Digital Object Identifier (DOI). Webpages for data download, documentation, and metadata are linked from each data set landing page. The data download page links to a zip file containing the data files for that data set in their original formats. The documentation webpage displays the titles to all five of the 2005 MA reports, with links to each of those reports. Each data set has

a full metadata record compliant with the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) schema, which can be displayed in various formats. The MA collection is available for free to all users from the SEDAC website, but users are required to log in using NASA's Earthdata login service in order to download data.

**SOCIOECONOMIC DATA AND APPLICATIONS CENTER (SEDAC)**  
*A Data Center in NASA's Earth Observing System Data and Information System (EOSDIS) — Hosted by CIRES at Columbia University*

Search SEDAC | Data | LOGIN

DATA | MAPS | THEMES | RESOURCES | SOCIAL MEDIA | ABOUT | HELP

### Millennium Ecosystem Assessment (MA)

Follow Us: [Twitter] [Facebook] [YouTube] [RSS] | Share: [Twitter] [Facebook]

**Collection Overview**

**Data Sets (6)**

- MA Population, v1 (1990–2002)

[+] Show All...

**MA Population, v1 (1990–2002)**

[Set Overview] [Data Download] [Documentation] [Metadata]

**Purpose:**

To preserve access to the original population data used by the Millennium Ecosystem Assessment (MA) and other related research.

**Abstract:**

The Millennium Ecosystem Assessment: MA Population provides data and information on baseline population as one of the drivers of ecosystem change. The data was used in estimating the magnitude of regional pressures on ecosystems. The MA Population data sets include Gridded Population of the World (GPW) Version 3, population grids from the Alpha version of the Global Rural-Urban Mapping Project (GRUMP), Global Subnational Infant Mortality Rates (Alpha version), and Global Subnational Prevalence of Child Malnutrition (Alpha version).

**Recommended Citation(s)\*:**

Millennium Ecosystem Assessment. 2005. Millennium Ecosystem Assessment: MA Population. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). <http://dx.doi.org/10.7927/H4CF9N1K>. Accessed DAY MONTH YEAR.

ENW (EndNote & RefWorks)†  
 RIS (Others)

\* When authors make use of data they should cite both the data set and the scientific publication, if available. Such a practice gives credit to data set producers and advances principles of transparency and reproducibility. Please visit the [data citations page](#) for details. Users who would like to choose to format the citation(s) for this dataset using a myriad of alternate styles can copy the DOI number and paste it into [Crosscite's website](#).

† For EndNote users, please check the Research Note field for issues with importing authors that are organizations when using the ENW file format.

**Available Formats:**

raster

## FIGURE 12.1

Landing page of the data set, Millennium Ecosystem Assessment: MA Population. Source: Millennium Ecosystem Assessment, Millennium Ecosystem Assessment: MA Population. (Palisades, NY: NASA Socioeconomic Data and Applications Center [SEDAC], 2005). doi:10.7927/H4CF9N1K.

## Lessons Learned

The MA collection data rescue experience is instructive for data producers who create or collect data, for repositories that acquire data from such projects, and for repositories that will be rescuing data. In addition to informing other data rescue efforts, the lessons of the MA collection data rescue also offer insight into potential risks of data management that can be mitigated by repositories through better coordination with data producers and anticipation of users' needs.

Clearly, data rescue would be much easier if data producers conducted due diligence during data creation and collection projects to ensure that all data produced has been properly prepared for preservation to enable its continuing use by others who are not part of the data study team. Such preparation should include the creation of complete documentation and provenance information. But in the absence of full documentation, even basic information on files, data sources, and the names and contact information of those involved would facilitate future preservation. Much time and effort can be wasted when such basic information is missing. Similarly, clearly identifying authorship and intellectual property rights is much more straightforward to do at the time when data is produced rather than years or decades later. Data repositories can provide guidance and tools for data producers to enhance their data documentation and provide users with more comprehensive information about the data, its collection, and its potential for use. Earlier involvement of data managers in national and international scientific research and assessment programs could also improve the development of appropriate data management policies, procedures, and incentives and increase the likelihood that resources would be allocated for their implementation.<sup>23</sup>

In many cases, research groups or assessment teams assemble data from multiple sources and integrate this data with their own, producing value-added data sets, models, or other research outputs. Again, clear documentation of these steps and careful attention to version control of both inputs and outputs are important in order to improve transparency and traceability of results. Such efforts are often neglected due to the assumption that input data is already sufficiently documented or due to time and resource limitations and competing priorities. More extensive use of workflow management tools and self-documenting data transformation and analysis packages may help address this problem in the future, as would publisher and funder requirements to deposit data in an approved archive in order to make data openly available.

Data repositories that acquire data from data producers and accept responsibility for the management of such resources need to discuss the opportunities for broad public dissemination with the data producers and come to an agreement regarding the expectations and responsibilities of both parties. As part of such negotiations with the data producers, the data repository should request and receive nonexclusive intellectual property rights that will allow anyone to archive, use,

integrate, and disseminate the data without restrictions, as long as attribution is provided for the source of the data. SEDAC tries to negotiate such unrestrictive rights for the data that it acquires so that the same rights can be offered to its users. These rights are described in each data set's online documentation and metadata.

It is also of course critical that data repositories take long-term data stewardship seriously, even if their primary focus is support for current data needs. They should attempt to develop appropriate preservation metadata in addition to discovery metadata for their holdings so that key information needed to understand and use data are not lost. Potential time-based dependencies should be identified to avoid losses due to media deterioration, technological obsolescence, or destruction schedules.<sup>24</sup> Information about the quality of the data and the results of data quality assessments should be accessioned with the data. Likewise, any rights agreement or other licenses obtained for the data should be archived. Repositories should manage their data holdings in accordance with the Open Archival Information Systems (OAIS) framework.<sup>25</sup> Data repositories need to conduct ongoing assessments of their data holdings to ensure that their data holdings have been properly prepared and effectively managed to enable usability by the communities served, even if the data is not planned for transfer to another facility. Plans for the sustainability or transition of the data infrastructure and holdings should be established by the repository so that access to the data can continue in the event of the termination of funding or operational authority of the repository. In the long run, it would be ideal for all data repositories to meet one or more standards for data stewardship, such as the Data Seal of Approval, the Trustworthy Repositories Audit & Certification (TRAC), or ISO 16363:2012, Space data and information transfer systems—Audit and certification of trustworthy digital repositories.<sup>26</sup> SEDAC has worked to meet the TRAC and ISO 16363:2012 standards, including a collaboration with the Columbia Libraries to ensure a long-term institutional home for all of SEDAC's data holdings.

Like the repositories that acquire data from producers, data repositories that engage in data rescue efforts need an established selection-and-appraisal process to select the data for curation and determine the appropriate level of service for continuing use of the data. A complete assessment of the candidate data rescue should be conducted to identify the effort and resources needed to meet basic preservation goals versus additional investments to meet current preservation and usability standards and expectations. When considering competing priorities for limited budgets, the potential value of scientific data to future scientific, historical, and policy research and applications should be considered both for data rescue and for current data management. Alternatively, it may be worth exploring whether members of the scientific community or another repository or entity might be able to contribute to or support the data rescue.

## Discussion and Conclusion

Unlike typical data curation efforts that are conducted at scientific data centers, data rescue may well require divergence from regular data curation procedures as tradeoffs may be necessary. The extent of such divergence may depend on the state of the data when it is acquired as well as on the availability of the data producers and data documentation. With the passage of time, the difficulty of any particular data rescue will inevitably increase, as data, documentation, and sources of information become more difficult if not impossible to access.

It is therefore important to move quickly when the need for a data rescue has been identified. In the case described here, SEDAC benefited from the relatively quick recognition of the need for a data rescue effort, that is, within one to two years of the NBII closure. However, the effort was also hampered by the poor state of the data more than seven years after the completion of the MA. Early identification of candidates for data rescue and the initiation of immediate action should increase the success of data rescue efforts. Similarly, the MA data rescue effort benefited from the familiarity that some SEDAC staff members had with the data being rescued. Such familiarity helped facilitate access to key scientists and critical information needed to document the data and determine access rights. Repositories, data centers, and archives that have worked with data that is at risk or with the associated scientific communities may be better positioned to take on data rescue activities in these areas.

## Acknowledgments

The work reported in this chapter was conducted with support from the National Aeronautics and Space Administration (NASA) under contract NNG13HQ04C for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC). We acknowledge the extensive efforts of many members of SEDAC's staff in carrying out the MA data rescue, especially that of information scientist Xiaoshi Xing.

## Notes

1. US Geological Survey, "NBII to Be Taken Offline Permanently in January," *USGS Access Newsletter* 14, no. 3 (Fall 2011), [http://www.usgs.gov/core\\_science\\_systems/Access/p1111-1.html](http://www.usgs.gov/core_science_systems/Access/p1111-1.html).
2. L. S. Tan, S. Burton, R. Crouthamel, A. van Engelen, R. Hutchinson, L. Nicodemus, T. C. Peterson, F. Rahimzadeh, *Guidelines on Climate Data Rescue*, WMO/TD No. 1210, ed. Paul Llansó and Hama Kontongomde (Geneva, Switzerland: World Meteorological Organization, 2004), <http://www.wmo.int/pages/prog/wcp/wcdmp/documents/WCDMP-55.pdf>.



3. Manola Brunet and Phil Jones, "Data Rescue Initiatives: Bringing Historical Climate Data into the 21st Century," *Climate Research* 47, no.1 (2011): 29, doi:10.3354/cr00960.
4. R. Elizabeth Griffin, "When Are Old Data New Data?" *GeoResJ* 6 (June 2015): 93, doi:10.1016/j.grj.2015.02.004.
5. F. Domínguez-Castro, Alexandre M. Ramos, Ricardo García-Herrera, and Ricardo M. Trigo, "Iberian Extreme Precipitation 1855/1856: An Analysis from Early Instrumental Observations and Documentary Sources," *International Journal of Climatology* 35, no. 1 (2015): 142–53, doi:10.1002/joc.3973.
6. M. Brunet, P. D. Jones, S. Jourdain, D. Efthymiadis, M. Kerrouche, and C. Boroneant, "Data Sources for Rescuing the Rich Heritage of Mediterranean Historical Surface Climate Data," *Geoscience Data Journal* 1, no. 1 (2014): 61–73, doi:10.1002/gdj3.4.
7. University of East Anglia (UEA) Climatic Research Unit, "Data," accessed November 23, 2015, <http://www.cru.uea.ac.uk/data>; International Research Institute for Climate and Society (IRI), "Climate Data Library," accessed November 23, 2015, <http://iri.columbia.edu/resources/data-library/>; US National Climatic Data Center, accessed November 23, 2015, <http://www.ncdc.noaa.gov>.
8. Intergovernmental Panel on Climate Change (IPCC), "Data Distribution Centre," accessed November 23, 2015, [http://www.ipcc-data.org/observ/clim/ar4\\_global.html](http://www.ipcc-data.org/observ/clim/ar4_global.html).
9. Paul N. Edwards, *A Vast Machine* (Cambridge, MA: MIT Press, 2010).
10. S. J. Hawkins, L. B. Firth, M. McHugh, E. S. Poloczanska, R. J. H. Herbert, M. T. Burrows, M. A. Kendall et al., "Data Rescue and Re-use: Recycling Old Information to Address New Policy Concerns," *Marine Policy* 42 (November 2013): 91–98, doi:10.1016/j.marpol.2013.02.001.
11. S. Levitus, "The UNESCO-IOC-IODE 'Global Oceanographic Data Archeology and Rescue' (GODAR) Project and 'World Ocean Database' Project," *Data Science Journal* 11 (2012): 46–71, doi:10.2481/dsj.012-014; Hawkins et al., "Data Rescue and Re-use"; Kenneth R. Knapp, John J. Bates, and Bruce Barkstrom, "Scientific Data Stewardship: Lessons Learned from a Satellite-Data Rescue Effort," *Bulletin of the American Meteorological Society* 88, no. 9 (2007): 1359–61, doi:10.1175/BAMS-88-9-1359; Patrick C. Caldwell, "Tide Gauge Data Rescue," *Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation* (Vancouver, British Columbia, Canada: UNESCO, 2013 ), 134–149. [http://cisra.org/docs/UNESCO\\_MOW2012\\_Proceedings\\_FINAL\\_ENG\\_Compressed.pdf](http://cisra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf); Jamus Collier, Stefanie Schumacher, Cornelia Behrens, Amelie Driemel, Michael Diepenbroek, Hannes Grobe, Taewoon Kim, Uwe Schindler, Rainer Sieger, and Hans-Joachim Wallrabe-Adams, "Rescued from the Deep: Publishing Scientific Ocean Drilling Long Tail Data," *GeoResJ* 6 (June 2015): 17–20, doi:10.1016/j.grj.2015.01.003.
12. David G. Gallaher, Garrett Campbell, Walter Meier, John Moses, and Dennis Wingo, "The Process of Bringing Dark Data to Light: The Rescue of the Early Nimbus Satellite Data," *GeoResJ* 6 (June 2015): 124–34, doi:10.1016/j.grj.2015.02.013.
13. Frank Kaspar, Birger Tinz, Hermann Mächel, and Lydia Gates, "Data Rescue of National and International Meteorological Observations at Deutscher Wetterdienst," *Advances in Science and Research* 12, no. 1 (2015): 57–61, doi:10.5194/asr-12-57-2015.
14. Karen M. Fallas, Robert B. MacNaughton, and Matthew J. Sommers, "Maximizing the Value of Historical Bedrock Field Observations: An Example from Northwest Canada," *GeoResJ* 6 (June 2015): 30–43, doi:10.1016/j.grj.2015.01.004.



15. Knapp, Bates, and Barkstrom, “Scientific Data Stewardship,” 1359.
16. Richard Munang, Johnson N. Nkem, and Zhen Han, “Using Data Digitalization to Inform Climate Change Adaptation Policy: Informing the Future Using the Present,” *Weather and Climate Extremes* 1 (September 2013): 17–18, doi:10.1016/j.wace.2013.07.001.
17. Lesley Wyborn, Leslie Hsu, Kerstin Lehnert, and Mark A. Parsons, “Guest Editorial: Special Issue: Rescuing Legacy Data for Future Science,” *GeoResJ* 6 (June 2015): 106–7, doi:10.1016/j.grj.2015.02.017.
18. NASA Socioeconomic Data and Applications Center, “Millennium Ecosystem Assessment (MA),” 2005, accessed September 26, 2015, <http://sedac.ciesin.columbia.edu/data/collection/ma>.
19. Millennium Ecosystem Assessment, *Ecosystems and Human Well-Being: Synthesis* (Washington DC: Island Press, 2005), <http://www.unep.org/maweb/documents/document.356.aspx.pdf>.
20. US Geological Survey, “NBII to Be Taken Offline.”
21. NASA Socioeconomic Data and Applications Center (SEDAC), “User Working Group,” accessed November 23, 2015, <http://sedac.ciesin.columbia.edu/user-working-group>.
22. NASA Socioeconomic Data and Applications Center (SEDAC), “Millennium Ecosystem Assessment (MA),” 2005, <http://sedac.ciesin.columbia.edu/data/collection/ma>.
23. Mustapha Mokrane and Mark A. Parsons “Learning from the International Polar Year to Build the Future of Polar Data Management,” *Data Science Journal* 13 (2014): PDA88–PDA93, doi:10.2481/dsj.IFPDA-15.
24. Robert R. Downs, “Data Rescue at a Scientific Data Center,” (presentation at the Best Practices Exchange [BPE] 2015, Harrisburg, PA, October 19–21, 2015), Columbia University Academic Commons, doi:10.7916/D8B857Q3.
25. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, Recommended Practice CCSDS 650.0-M-2, Magenta Book, Issue 2 (Washington, DC: CCSDS Secretariat, June 2012), <http://public.ccsds.org/publications/archive/650x0m2.pdf>.
26. Data Seal of Approval homepage, accessed November 23, 2015, <http://datasealofapproval.org/en/>; Center for Research Libraries, “TRAC Metrics,” accessed November 23, 2015, <http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac>; International Organization for Standardization, “ISO 16363:2012: Space Data and Information Transfer Systems—Audit and Certification of Trustworthy Digital Repositories,” February 15, 2012, accessed November 23, 2015, [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=56510](http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510).

## Bibliography

- Brunet, M., P. D. Jones, S. Jourdain, D. Efthymiadis, M. Kerrouche, and C. Boroneant. “Data Sources for Rescuing the Rich Heritage of Mediterranean Historical Surface Climate Data.” *Geoscience Data Journal* 1, no. 1 (2014): 61–73. doi:10.1002/gdj3.4.
- Brunet, Manola, and Phil Jones. “Data Rescue Initiatives: Bringing Historical Climate Data into the 21st Century.” *Climate Research*, 47, no.1 (2011): 29–40. doi:10.3354/cr00960.

- Caldwell, Patrick C. "Tide Gauge Data Rescue." *Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation*, 134–149. Vancouver, British Columbia, Canada: UNESCO 2013, 2012. [http://cisra.org/docs/UNESCO\\_MOW2012\\_Proceedings\\_FINAL\\_ENG\\_Compressed.pdf](http://cisra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf).
- Center for Research Libraries, "TRAC Metrics," accessed November 23, 2015, <http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac>.
- Collier, Jamus, Stefanie Schumacher, Cornelia Behrens, Amelie Driemel, Michael Diepenbroek, Hannes Grobe, Taewoon Kim, Uwe Schindler, Rainer Sieger, and Hans-Joachim Wallrabe-Adams. "Rescued from the Deep: Publishing Scientific Ocean Drilling Long Tail Data." *GeoResJ* 6 (June 2015): 17–20. doi:10.1016/j.grj.2015.01.003.
- Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. Recommended Practice, CCSDS 650.0-M-2, Magenta Book, Issue 2. Washington, DC: CCSDS Secretariat, June 2012. <http://public.ccsds.org/publications/archive/650x0m2.pdf>.
- Data Seal of Approval homepage, accessed November 23, 2015, <http://datasealofapproval.org/en/>.
- Domínguez-Castro, F., Alexandre M. Ramos, Ricardo García-Herrera, and Ricardo M. Trigo. "Iberian Extreme Precipitation 1855/1856: An Analysis from Early Instrumental Observations and Documentary Sources." *International Journal of Climatology* 35, no. 1 (2015): 142–53. doi:10.1002/joc.3973.
- Downs, Robert R. "Data Rescue at a Scientific Data Center." Presentation, Best Practices Exchange (BPE) 2015, Harrisburg, PA, October 19–21, 2015. Columbia University Academic Commons, doi:10.7916/D8B857Q3.
- Edwards, Paul N. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press, 2010.
- Fallas, Karen M., Robert B. MacNaughton, and Matthew J. Sommers. "Maximizing the Value of Historical Bedrock Field Observations: An Example from Northwest Canada." *GeoResJ* 6 (June 2015): 30–43. doi:10.1016/j.grj.2015.01.004.
- Gallaher, David, G. Garrett Campbell, Walter Meier, John Moses, and Dennis Wingo. "The Process of Bringing Dark Data to Light: The Rescue of the Early Nimbus Satellite Data." *GeoResJ* 6 (June 2015): 124–34. doi:10.1016/j.grj.2015.02.013.
- Griffin, R. Elizabeth. "When Are Old Data New Data?" *GeoResJ* 6 (June 2015): 92–97. doi:10.1016/j.grj.2015.02.004.
- Hawkins, S. J., L. B. Firth, M. McHugh, E. S. Poloczanska, R. J. H. Herbert, M. T. Burrows, M. A. Kendall et al. "Data Rescue and Re-use: Recycling Old Information to Address New Policy Concerns." *Marine Policy* 42 (November 2013): 91–98. doi:10.1016/j.marpol.2013.02.001.
- Intergovernmental Panel on Climate Change (IPCC), "Data Distribution Centre," accessed November 23, 2015, [http://www.ipcc-data.org/observ/clim/ar4\\_global.html](http://www.ipcc-data.org/observ/clim/ar4_global.html).
- International Organization for Standardization, "ISO 16363:2012: Space Data and Information Transfer Systems—Audit and Certification of Trustworthy Digital Repositories," February 15, 2012, accessed November 23, 2015, [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=56510](http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510).
- International Research Institute for Climate and Society (IRI), "Climate Data Library," accessed November 23, 2015, <http://iri.columbia.edu/resources/data-library/>.

- Kaspar, Frank, Birger Tinz, Hermann Mächel, and Lydia Gates. “Data Rescue of National and International Meteorological Observations at Deutscher Wetterdienst.” *Advances in Science and Research* 12, no. 1 (2015): 57–61. doi:10.5194/asr-12-57-2015.
- Knapp, Kenneth R., John J. Bates, and Bruce Barkstrom. “Scientific Data Stewardship: Lessons Learned from a Satellite-Data Rescue Effort.” *Bulletin of the American Meteorological Society* 88, no. 9 (2007): 1359–61. doi:10.1175/BAMS-88-9-1359.
- Levitus, S. “The UNESCO-IOC-IODE ‘Global Oceanographic Data Archeology and Rescue’ (GODAR) Project and ‘World Ocean Database’ Project.” *Data Science Journal* 11 (2012): 46–71. doi:10.2481/dsj.012-014.
- Millennium Ecosystem Assessment. *Ecosystems and Human Well-Being: Synthesis*. Washington DC: Island Press, 2005. <http://www.unep.org/maweb/documents/document.356.aspx.pdf>.
- Mokrane, Mustapha, and Mark A. Parsons. “Learning from the International Polar Year to Build the Future of Polar Data Management.” *Data Science Journal* 13 (2014): PDA88–PDA93. doi:10.2481/dsj.IFPDA-15.
- Munang, Richard, Johnson N. Nkem, and Zhen Han. “Using Data Digitalization to Inform Climate Change Adaptation Policy: Informing the Future Using the Present.” *Weather and Climate Extremes* 1 (September 2013): 17–18. doi:10.1016/j.wace.2013.07.001.
- NASA Socioeconomic Data and Applications Center (SEDAC). “Millennium Ecosystem Assessment (MA).” 2005. <http://sedac.ciesin.columbia.edu/data/collection/ma>.
- . “Millennium Ecosystem Assessment: MA Population.” 2005. doi:10.7927/H4CF9N1K.
- . “User Working Group,” accessed November 23, 2015, <http://sedac.ciesin.columbia.edu/user-working-group>.
- Tan, L. S., S. Burton, R. Crouthamel, A. van Engelen, R. Hutchinson, L. Nicodemus, T. C. Peterson, F. Rahimzadeh. *Guidelines on Climate Data Rescue*. WMO/TD No. 1210. Edited by Paul Llansó and Hama Kontongomde. Geneva, Switzerland: World Meteorological Organization, 2004. <http://www.wmo.int/pages/prog/wcp/wcdmp/documents/WCDMP-55.pdf>.
- US National Climatic Data Center, accessed November 23, 2015, <http://www.ncdc.noaa.gov>.
- University of East Anglia (UEA) Climatic Research Unit, “Data,” accessed November 23, 2015, <http://www.cru.uea.ac.uk/data>.
- US Geological Survey. “NBII to Be Taken Offline Permanently in January.” *USGS Access Newsletter* 14, no. 3 (Fall 2011). [http://www.usgs.gov/core\\_science\\_systems/Access/p1111-1.html](http://www.usgs.gov/core_science_systems/Access/p1111-1.html).
- Wyborn, Lesley, Leslie Hsu, Kerstin Lehnert, and Mark A. Parsons. “Guest Editorial: Special Issue: Rescuing Legacy Data for Future Science.” *GeoResJ* 6 (June 2015): 106–7. doi:10.1016/j.grj.2015.02.017.



## BIOGRAPHIES

# Contributor Biographies

## Editor Biography

**Lisa R. Johnston** is an Associate Librarian at the University of Minnesota, Twin Cities. Johnston serves as the libraries' Research Data Management/Curation Lead and as Co-Director of the University Digital Conservancy, the University of Minnesota's institutional repository. In 2014, Johnston led the team that developed and launched the Data Repository for the University of Minnesota (DRUM), <http://hdl.handle.net/11299/166578>. She serves as principal investigator of the multi-institution collaboration, the Data Curation Network project, which launched in 2016 with funding from the Alfred P. Sloan Foundation. Johnston has presented internationally on topics of academic library services for research data management, authored research articles on data management topics, and co-edited the book *Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers* (Purdue University Press, eds. Carlson and Johnston, 2015), which details a variety of educational approaches used in data management training for STEM graduate students. Prior to becoming a librarian, Johnston was a science writer and assistant editor for *Sky & Telescope* magazine. Johnston holds a master of library science and bachelor of science in astrophysics, both from Indiana University, and was certified by the Society of American Archivists as a Digital Curation Specialist. Her ORCID is <http://orcid.org/0000-0001-6908-9240>.

## Author Biographies

**Karen S. Baker**, after careers in oceanography and data management, is a doctoral student at University of Illinois Urbana-Champaign in the School of Information Sciences. Having worked as an information manager with the Long-Term Ecological Research program, her work with data stretches from bio-optics and

field science to sociotechnical systems and the information sciences. Karen's interests are in study of the data ecosystem and how data practices can inform the growth of information infrastructures that support research. Her studies explore the continuing development of data repositories and information environments that can in turn facilitate collaborative work and collective learning.

**Eugene Barsky** is Research Data Librarian at the UBC Library. He received his MLIS from the University of British Columbia in 2005. His recent peer recognition included American Society for Engineering Education and Special Library Association awards. He published more than twenty peer-reviewed papers and presented at more than forty conferences. Eugene is an adjunct faculty member at the iSchool at UBC, teaching courses in science librarianship and research data management, and is an active member of the Pacific Northwest data curators group.

**Kristin Briney** is the Data Services Librarian at the University of Wisconsin-Milwaukee, where she advises researchers on data management plans and data best practices. She has a PhD in chemistry and a master's degree in library and information studies. Kristin is the author of the book *Data Management for Researchers* and blogs about how to manage research data at <http://dataabinitio.com>.

**Robert S. Chen**, PhD, is director and senior research scientist at CIESIN, the Center for International Earth Science Information Network at Columbia University's Earth Institute in New York. He manages the NASA Socioeconomic Data and Applications Center (SEDAC), part of NASA's network of earth science data centers. He co-manages the Intergovernmental Panel on Climate Change (IPCC) Data Distribution Center and is a co-chair of the Thematic Network on Sustainable Development Data of the United Nations Sustainable Development Solutions Network and the Data Sharing Working Group of the Group on Earth Observations. He is a member of the Governing Council of the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan and the Council of the American Geographical Society. He received his PhD in geography from the University of North Carolina at Chapel Hill and holds BS and MS degrees from the Massachusetts Institute of Technology.

**Robert R. Downs**, PhD, is a senior staff associate officer of research and serves as senior digital archivist and acting head of cyberinfrastructure and informatics research and development at CIESIN, the Center for International Earth Science Information Network, a research and data center of the Earth Institute of Columbia University. He is Vice-Chair of the Columbia University Morningside Institutional Review Board and holds a PhD in information management from the Stevens Institute of Technology. He is a member of the board of directors of the Foundation for Earth Science Information Partners (ESIP) and is a member

of the editorial board of the *Data Science Journal*. He has been a Senior Member of the Association for Computing Machinery (ACM) since 2012 and is a member of the American Geophysical Union (AGU), the Association for Information Science and Technology (ASIS&T), and the International Association for Social Sciences Information Services and Technology (IASSIST).

**Ruth E. Duerr** is a generalist with interests in nearly everything from music and art to science, data, and policy with stops for many other things along the way. She has been actively involved in earth and space sciences together with the information and data sciences for many years. She has a passion for encouraging researchers to become better data managers, demonstrated in part through teaching courses, developing and presenting a variety of workshops, and editing of the Federation of Earth Science Information Partners (ESIP) peer-reviewed Data Management short course (learning modules that are freely available online for self-paced learning). Her lectures at meetings and conferences introduce audiences (from various disciplines in earth and space sciences, social sciences, and computing and information sciences) to data management issues and solutions. She also teaches data curation at the School of Information Sciences at the University of Illinois at Urbana Champaign, contributing to the expertise of new data practitioners.

**Ixchel M. Faniel**, PhD is a Research Scientist at OCLC. Her interests include improving how people discover, access, and use and reuse content. Her current research examines how academics manage, share, and reuse research data and librarians' experiences designing and delivering supportive research data services. Her research has been funded by the National Science Foundation, Institute of Museum and Library Services, and National Endowment for the Humanities. Her ORCID is <http://orcid.org/0000-0001-7302-5936>.

**Kathleen Fear** is the data librarian at the University of Rochester's River Campus Libraries and leads the Numeric, Spatial and Research Data Services team in developing services for managing and sharing research data. She holds a PhD from the University of Michigan's School of Information.

**Katherine J. Gerwig** is an Information Commons Specialist at Metropolitan State University in Minnesota. She recently received an MA in library and information studies from the University of Wisconsin–Madison and completed a practicum placement with Research Data Services at the University of Minnesota.

**Abigail Goben** is an Information Services and Liaison Librarian at the University of Illinois at Chicago Library of the Health Sciences. She is liaison to the College of Dentistry and collaborates with health science researchers on research data

management. Her research focuses on institutional data policy, self-education in research data management, and open access. She blogs at <http://hedgehoglibrarian/>.

**Heidi J. Imker** is the Director of the Research Data Service (RDS) at the University of Illinois at Urbana-Champaign. The RDS is a campus-wide service headquartered in the University Library that provides the Illinois research community with the expertise, tools, and infrastructure necessary to manage and steward research data. Prior to joining the library, Heidi was the Executive Director of the Enzyme Function Initiative, a large-scale collaborative center involving nine universities, funded by the National Institutes of Health, and located in the Institute for Genomic Biology. Heidi holds a PhD in biochemistry from the University of Illinois and did her postdoctoral research at the Harvard Medical School.

**Mayu Ishida** is a science reference librarian at the University of British Columbia (UBC), and her research interests include data management practices and needs. Before joining UBC, she was the research services librarian at the University of Manitoba and facilitated services in data management and open access. She completed her master of library and information studies at UBC and worked on the Canadian International Polar Year data curation project at the University of Alberta Libraries.

**Christine Kollen** is the Data Curation Librarian at the University of Arizona Libraries in the Office of Digital Innovation and Stewardship. She leads the University of Arizona's efforts in providing data management support for researchers and graduate students. She is chair of the Campus Data Management and Curation Committee and the project manager for the Data Management and Data Curation pilot project. She also leads the library's GIS and geospatial data services and is the project manager for the Spatial Data Explorer, the library's geospatial data portal. She co-presented "Developing Research Data Services Vision(s): An Analysis of North American Academic Libraries" at the 2015 IASSIST conference and has also presented and written on developing geoportals and providing user-friendly access to geospatial and historic census data collections.

**Inna Kouper** is a research scientist and assistant director of the Data to Insight Center at Indiana University Bloomington. Her research interests are in the history and sociology of knowledge production and dissemination, with a particular emphasis on the sociotechnical (STS) approaches to emerging technologies and data practices. Dr. Kouper has a PhD in information science from Indiana University Bloomington and a PhD in sociology from the Institute of Sociology, Russian Academy of Sciences, Moscow, Russia.



**Larry Laliberté** is a librarian with over ten years' experience working with GIS and spatial data. Currently he is the Geospatial Data Services Librarian at the University of Alberta, where much of his work revolves around analyzing and synthesizing spatial information at many scales, across many disciplines, in various formats. Over the last decade, he has developed and maintained an online collection of historical maps of Manitoba and recently taken a great interest in developing best practices for the long-term preservation of digital geospatial data.

**Amber Leahey** is the Data and Geospatial Librarian at Scholars Portal, the digital library project of the Ontario Council of University Libraries. She supports digital data services within OCUL, including <odesi>, Scholars GeoPortal, and Scholars Portal's Dataverse Network. Previously, she was the Data Services Metadata Librarian at Scholars Portal. Amber is actively involved in a variety of Canadian and international library research data communities, including IASSIST, Data Liberation Initiative (DLI), Data Documentation Initiative (DDI), Research Data Canada (RDC), and the Canadian Association of Research Libraries (CARL) Portage Network.

**Natsuko Nicholls**, PhD, is a Data Manager & Analyst for the IRIS program (Institute for Research on Innovation and Science) at the University of Michigan's Institute for Social Research (ISR). Until recently, she was the Research Data Consultant and Assistant Professor in Virginia Tech University Libraries. She dedicated her efforts to developing and implementing the libraries' research data services with a focus on data consulting. Her data management consulting activities ranging from direct assistance with faculty in writing data management plans to indirect assistance in developing data management workflows. She also offered instruction on data management concepts, workflows, and best practices for diverse audiences comprised of first-year undergraduate students to early-career and senior researchers.

**Karl Nilsen** was the Research Data Librarian at the University of Maryland Libraries, College Park. In this role, he provided faculty and students with advice and assistance on all aspects of data management and dissemination. In addition, he worked on knowledge management, digital curation, and scholarly communication activities as a member of the libraries' Digital Programs and Initiatives team.

**Andrea L. Ogier**, MLIS, is Associate Director of Data Services, and Assistant Professor in the University Libraries at Virginia Tech. She leads the Data Services Unit and works to build library services around data and informatics consulting, data management, data curation and data literacy, and to integrate those services into the university's research enterprise.

**Ryan Speer**, MLIS, is University Records Manager and Assistant Professor in Virginia Tech University Libraries. His research interests include archival replevin and other topics associated with public records management.

**Leanne Trimble** has recently joined the University of Toronto Libraries (UTL) as Data & Statistics Librarian. She supports the university community in the discovery, creation, and use of statistics and numeric data. This includes participating on UTL's Research Data Management Working Group. Prior to joining UTL, Leanne worked for Scholars Portal (Ontario Council of University Libraries), where she coordinated data services including <odesi>, Scholars GeoPortal, and the Scholars Portal Dataverse Network.

**Cynthia R. Hudson Vitale** is the Data Services Coordinator in Data & GIS Services at Washington University in St. Louis Libraries. In this position, Cynthia leads research data services and curation efforts for the libraries. Since coming into this role in 2012, she has worked on faculty projects to facilitate data sharing and interoperability while meeting faculty research data needs throughout the research life cycle. She has also worked across the university to improve research reproducibility, addressing both technical and cultural barriers. She currently serves as the Visiting Program Officer for SHARE with the Association of Research Libraries.

**Jon Wheeler** is a Data Curation Librarian within the University of New Mexico's College of University Libraries and Learning Sciences. As a member of the libraries' Research Data Services program, he has a principal focus on the development of research data ingest, packaging, and archiving workflows, which facilitate preservation and compliance with funder requirements. His ORCID is <http://orcid.org/0000-0002-7166-3587>.

**Sarah C. Williams** is the Life Sciences Data Services Librarian at the University of Illinois at Urbana-Champaign. Focusing on the research data needs of life scientists on campus, she conducts training, provides individualized consultations, reviews data management plans, and develops web resources. Her research concentrates on data practices in the life sciences and services that can facilitate better data practices. She has a bachelor's degree in soil and crop science from Purdue University, an MLS from Indiana University, and a master's degree in information systems from Illinois State University.

**Elizabeth Yakel**, PhD, is a Professor at the University of Michigan School of Information, where she teaches in the archives and records management and digital preservation areas. Her research focuses on users of primary sources, particularly how to facilitate access to digital archives and the reuse of research data. She is currently working on an IMLS-funded project, "Qualitative Data Reuse: Records

of Practice in Educational Research and Teacher Development,” which examines data reuse by researchers and teacher-educators of digital video of classroom activities (<http://qualitativedatareuse.org>). Her ORCID is <http://orcid.org/0000-0002-8792-6900>.

**Lisa D. Zilinski** is the Research Data Consultant at Carnegie Mellon University. As part of the Research Curation Division, she consults and collaborates with faculty, staff, and students to identify data literacy opportunities, develops learning plans and tools for data education, and investigates and develops programmatic and sustainable data services for the libraries. Her research interests include data informed learning, data management principles, data policy, and information dissemination and access practice.