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

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).[1] 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.[3] 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[4]

Sometimes primarily associated with the sciences, data can be found in any discipline and in many forms.[5] 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.[6] 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.[7] 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.[8]

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."[9] 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.[10]

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.[11]

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 lifecycle 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.[12]

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.[13] 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.[14] 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."[15] 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.[16] 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.[17] For example, a key turning point in the US was the February 22, 2013 memorandum[18] 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[19]

Three years after the OSTP directive, "policies to make data and publications resulting from federally funded research publicly accessible are becoming the norm."[20] 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.[21] 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.[22] 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.[23]

Complex? Absolutely. **Fortunately, Chapter 2, titled "Institutional, Funder, and Journal Data Policies" by Kristin Briney, Abigail Goben, 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.[24] This trend continues today and sources such as Retraction Watch regularly report examples of publishers responding to data-related issues in publications.[25] 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.[26] 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.[27] 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."[28] Indeed, one such retraction occurred in 2015, albeit in a different journal (*Frontiers in Neuroscience*), due to an author refusing to share their data.[29]

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 PANGEA), 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.[30] 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."[31] 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.[32] 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."[33] 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."[34] 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.[35] 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."[36] The European Commission has established a pilot program for data sharing through its Horizon 2020 granting arm.[37] 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."[38] **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).[39] 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.[40]

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.[41] As one blogger quips, "'Send me your data—pdf is fine,' said no one ever."[42] 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."[43] Other widespread data sharing approaches, such as posting data to a project website or sharing data upon request, may not 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.[44] Yet, data sharing is still happening and data curation efforts may help mitigate these error prone approaches. Imkers' 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.[45] 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.[46] **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.[47] 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.

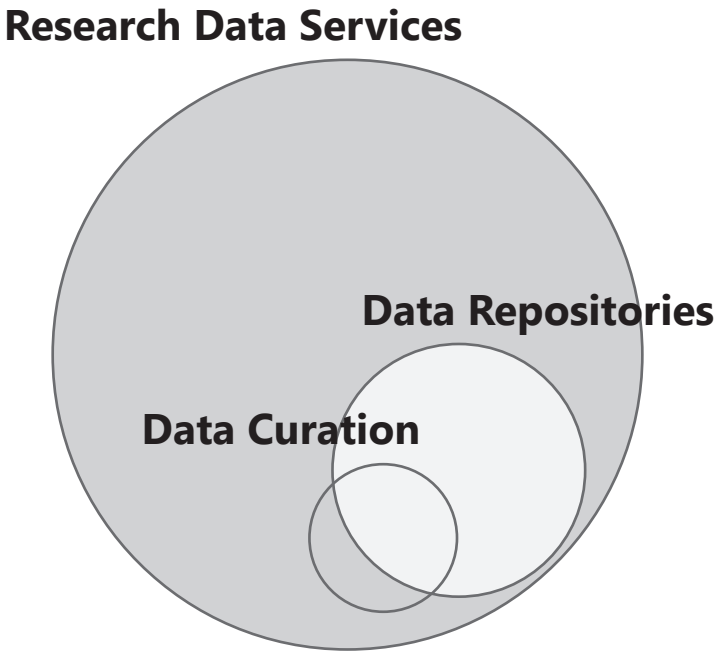

FIGURE INTRO.1
Data curation as a subset of research data services. Note that data curation services may support or overlap with local data repository services, or curation services may be provided for data that are deposited elsewhere, such as disciplinary repositories or non-accessible (dark) storage.

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.[48] 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.[49] Their in-depth resources help researchers better prepare their data for eventual deposit into the DataOne connected archives.[50] 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.[51]

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.[52] 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.[53] 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."[54] **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.[55]

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.[56] 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.[57] 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.[58] 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.[59]

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."[60] They go on to provide 10 ways to "distinguish scrutiny from harassment."[61] 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.'[62]

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.[63] 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.[64] 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…"[65] 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.[66] 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.[67] 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."[68] 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.[69]

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

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.[71]

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.[72] 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.[73] 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.
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.
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 August18, 2015, 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.
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.
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.

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.
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.
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; 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 polices 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.
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.
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.
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.

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.
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.
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. Fridsma, "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/callee/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.
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.
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/grbcat/grbcat.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.

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.

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.

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 August18, 2015. 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.

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.

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.

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.

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.

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/grbcat/grbcat.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.

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.

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 Vie*w. 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.

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.

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.

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. Feburary 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. 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.

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.

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/callee/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.

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.

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.

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.