U.S. Copyright Law (title 17 of U.S. code) governs the reproduction and redistribution of copyrighted material.

Downloading this document for the purpose of redistribution is prohibited.
Research Data Management
Practical Strategies for Information Professionals

Edited by Joyce M. Ray

Charleston Insights in Library, Archival, and Information Sciences

Purdue University Press
West Lafayette, Indiana
Contents

Introduction to Research Data Management 1
Joyce M. Ray

PART 1: UNDERSTANDING THE POLICY CONTEXT

1 The Policy and Institutional Framework 25
James L. Mullins

2 Data Governance: Where Technology and Policy Collide 45
MacKenzie Smith

PART 2: PLANNING FOR DATA MANAGEMENT

3 The Use of Life Cycle Models in Developing and Supporting Data Services 63
Jake Carlson

4 Data Management Assessment and Planning Tools 87
Andrew Sallans and Sherry Lake

5 Trustworthy Data Repositories: The Value and Benefits of Auditing and Certification 109
Bernard F. Reilly, Jr., and Marie E. Waltz
PART 3: MANAGING PROJECT DATA

6 Copyright, Open Data, and the Availability-Usability Gap: Challenges, Opportunities, and Approaches for Libraries 129
    Melissa Levine

7 Metadata Services 149
    Jenn Riley

8 Data Citation: Principles and Practice 167
    Jan Brase, Yvonne Socha, Sarah Callaghan, Christine L. Borgman, Paul F. Uhlir, and Bonnie Carroll

PART 4: ARCHIVING AND MANAGING RESEARCH DATA IN REPOSITORIES

9 Assimilating Digital Repositories Into the Active Research Process 189
    Tyler Walters

10 Partnering to Curate and Archive Social Science Data 203
    Jared Lyle, George Alter, and Ann Green

11 Managing and Archiving Research Data: Local Repository and Cloud-Based Practices 223
    Michele Kimpton and Carol Minton Morris

12 Chronopolis Repository Services 239
    David Minor, Brian E. C. Schottlaender, and Ardys Kozbial

PART 5: MEASURING SUCCESS

13 Evaluating a Complex Project: DataONE 255
    Suzie Allard
14 What to Measure? Toward Metrics for Research Data Management  
Angus Whyte, Laura Molloy, Neil Beagrie, and John Houghton  

PART 6: BRINGING IT ALL TOGETHER: CASE STUDIES  

15 An Institutional Perspective on Data Curation Services: A View from Cornell University  
Gail Steinhart  

16 Purdue University Research Repository: Collaborations in Data Management  
D. Scott Brandt  

17 Data Curation for the Humanities: Perspectives From Rice University  
Geneva Henry  

18 Developing Data Management Services for Researchers at the University of Oregon  
Brian Westra  

CLOSING REFLECTIONS: LOOKING AHEAD  

19 The Next Generation of Challenges in the Curation of Scholarly Data  
Clifford Lynch  

Index
Interest in research data has grown substantially over the past decade. The reason for this is evident: the digital revolution has made it far easier to store, share, and reuse data. Scientific research data are now almost universally created and collected in digital form, often in staggering quantities, and all disciplines are making increasing use of digital data. Data sharing increases the return on the large investments being made in research and has the potential to exponentially advance human knowledge, promote economic development, and serve the public good, all while reducing costly data duplication.

The Human Genome Project is a well-known example of the return on public investment resulting from collaborative research and data sharing. The project began in 1990 as an international effort to identify and map the sequence of the more than 20,000 genes of the human genome and to determine the sequence of chemical base pairs that make up DNA. Completed in 2003, the project produced GenBank, a distributed database that stores the sequence of the DNA in various locations around the world. The data are publicly accessible and continue to be mined for research in fields from molecular medicine and biotechnology to evolution. Findings have led to the development of genetic tests for predisposition to some diseases, and ongoing research is investigating potential disease treatments. GenBank now supports a multibillion-dollar genomics research industry to develop DNA-based products.

The success of GenBank and other highly visible research projects has drawn the attention of national governments and international organizations
to the potential of data sharing and international collaboration to solve some of the grand challenges facing the world today, from disease prevention and treatment to space exploration and climate change.

But having an interest in data sharing is only the first step in doing it successfully. In order for data to be shared among research teams and maintained for reuse over long periods of time, another grand challenge must be solved—preserving all this digital data and managing it so that it can be stored efficiently, discovered by secondary users, and used with confidence in its authenticity and integrity. When datasets were shared only among colleagues known to each other, trust was implicit. If data are to be made widely available and used by people with no personal knowledge of their creators, and for different purposes than those for which they were created, then trust must derive from how the data are managed and documented.

Required documentation includes not only search terms for future data discovery (descriptive metadata), but also evidence of the data’s provenance (how, when, where, why, and by whom it was created), its chain of custody, and information on how it has been managed to mitigate the risk of data loss or corruption. This is true for the “big data” projects that have captured the attention of the news media, and it is just as true and even

<table>
<thead>
<tr>
<th>Total Grants over $500</th>
<th>12,025</th>
<th>$2,865,388,605</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number Grants</strong></td>
<td>2404</td>
<td>9621</td>
</tr>
<tr>
<td><strong>Total Dollars</strong></td>
<td>$1,747,957,7451</td>
<td>$1,117,431,154</td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>$38,131,952- $300,000</td>
<td>$300,000- $579</td>
</tr>
<tr>
<td><strong>Number of grants</strong></td>
<td>254</td>
<td>11,771</td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>$38,131,952- 1,034,150</td>
<td>1,029,9984- $579</td>
</tr>
</tbody>
</table>

Table 1. NSF 2007 award distribution by award size. Courtesy of Bryan Heidorn.
more challenging for the smaller projects that account for the majority of research grants awarded by the National Science Foundation (NSF).

In Table 1, Bryan L. Heidorn demonstrates that the top 20 percent of NSF grants awarded in 2007 accounted for just over 50 percent of total funds spent, and the top 254 grants (2 percent) received 20 percent of the total. The remaining funds were distributed among 11,771 grants in amounts ranging from just over $500 to more than $1,000,000, with the average award in the range of $200,000 (Heidorn, 2008). Heidorn argues that the data in the top 20 percent of awards are more likely to be well curated than data in the 80 percent generated by smaller grants, and that it is important to improve data management practices in these smaller projects in order to maximize the return on investment.

Data that result from smaller projects often are more difficult to manage than big data because they are highly heterogeneous, require more individual attention per byte, and tend to be less well documented. Academic libraries generally lack the capacity to manage the large volumes associated with big data, but they may be well equipped to assist with managing smaller data projects. For example, they may recommend sustainable file formats and file organization, advise on intellectual property issues for data reuse, assist with determining appropriate metadata and data citation practices, and provide repository services for managing current research data as well as for archiving of data after project completion.

The volume of research data began to accumulate in very large quantities in the 1990s. Recognition that long-term maintenance of digital data requires an investment in human capital and infrastructure has grown over the past 30 years, but at a slower pace than the data itself. Federally funded research on digital libraries began in 1994, with six grants awarded in the NSF's Digital Libraries Initiative I. However, interest in digital preservation and best practices for the long-term management of digital data lagged behind research on digital library development. This was due in part to a false sense of confidence, based on ever-declining data storage costs and the belief that improvements in search algorithms would eliminate the need for concerns about such mundane topics as data organization, descriptive metadata, and file management. Federal funding for the applied research necessary to develop models and protocols for digital preservation and data management has been far more
modest than funding for the basic research that is at the heart of the NSF's mission.

Fortunately, the library and archival communities, with their long experience with information organization and documentation, have become deeply involved in the development of principles and best practices for managing digital data for long-term use. These principles and protocols now are being implemented as services, exemplified by the essays and case studies in this volume. While much of the work to develop implementation strategies for curating research data has taken place in research universities, largely in the scientific disciplines, the principles, practices, tools, and services described here have broad implications for all disciplines and all organizations with a preservation mission.

THE ARCHIVAL PERSPECTIVE

Archivists are responsible for preserving records, that is, the documentation of activities of the organization within which an archive is located. An organization's records provide evidence of its activities and policies, as well as information resulting from those activities. In order to serve as evidence, or proof, records must have authenticity, inferred by documentation of an unbroken chain of physical custody. They must also have integrity, showing that they have not been corrupted, and that any alterations have been authorized and documented to show what changes were made, when, why, and by whom.

Digital preservation activities, however, are likely to result in some alteration of the original digital object over time, in the course of migration or other preservation action. For example, even the simple act of opening a digital file automatically changes the "last modified" date and decreases the evidence of its integrity. How much alteration is acceptable? Can documentation about alterations compensate for the inability to preserve the exact form of the original for reuse? If so, what kinds of documentation are needed? Secondary users want access to a wide range of digital content, but in order for that information to have continuing value for scholarly research and to provide evidence for purposes of accountability or legal standing if required, the preserved data must include contextual information not only about the circumstances of its creation, but also about how it has been managed over time. Data repositories that aspire to trustworthiness must in-
clude documentation of all "events" that result in any changes to the digital objects they contain in the course of their ongoing preservation activities.

**HISTORICAL BACKGROUND: DATA AS EVIDENCE**

The science of diplomastics, which has guided the development of archival science, originated in the 17th century from the same need to create trusted documentation about events and transactions that now informs criteria for the management and evaluation of digital repositories to assess their trustworthiness. As government and commerce expanded over larger territories, states and merchants could no longer deal directly with the people they governed and with whom they conducted business. Therefore, they needed to create more documentation of transactions than had previously been required. Diplomastics provides the theoretical framework for a system of recordkeeping to verify and organize information that can be recognized as trustworthy (Gilliland-Swetland, 2000).

One of the fundamental principles of diplomastics is *provenance*, which documents the origin, lineage, or pedigree of an information object. Provenance is central to the ability to validate, verify, and contextualize digital objects, and it provides a large part of the context of meaning of an information object. It is vital for assessing the source, authority, accuracy, and value of the information contained in that object.

Digital preservation aims to ensure the maintenance over time of the value of digital objects. The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Authenticity Task Force, led by Luciana Duranti at the University of British Columbia, observed that users want to know that digital objects are what they purport to be (they are authentic), and that they are complete and have not been altered or corrupted (they have integrity) (Gilliland-Swetland and Eppard, 2000). Documents that lack authenticity and integrity have limited value as evidence or even as citable sources of information. And because digital objects are more susceptible to alteration and corruption than paper records, extra care must be taken to establish the authenticity and trustworthiness of digital objects.

Most online users begin with a presumption of authenticity, unless some concern arises that causes them to question it, but this may be changing in the digital environment as more challenges to the authenticity of data arise from charges of plagiarism, faulty research methods, and
even outright fraud. The only way users who do not have direct knowledge of an object's origin and management can trust its authenticity is for the organization that has taken custody of it to adequately and transparently document the provenance and process of ingest (acquisition or deposit into a digital repository), as well as its management within the repository.

The Research Roadmap Working Group of DigitalPreservationEurope (2007) identified five levels of preservation:

1. Digital object level, associated with issues of migration/emulation, experimentation, and acceptable loss;
2. Collection level, associated with issues of interoperability, metadata, and standardization;
3. Repository level, including policies and procedures;
4. Process level, associated with issues of automation and workflow; and
5. Organizational level, including issues of governance and sustainability.

All of these levels should be considered in designing data services. In order to share datasets across a wide variety of disciplines with different research interests, protocols must be established for describing and documenting data consistently. As repositories move from in-house operations to core services, they become an essential part of the digital infrastructure and must meet high standards of trustworthiness. Decisions made early in the data creation and active management phases of a research project inevitably affect how well the data can later be documented, preserved, and reused, so long-term preservation should be considered early in the planning process.

THE LIBRARY PERSPECTIVE

While it is an oversimplification to say that archives are about preservation and libraries are about access, it is fair to say that the most valuable contribution of archives to the digital infrastructure has been the principle of context for future use through data documentation and rules of evidence. The greatest contribution of libraries is most likely their emphasis on services, providing the basis not only for future access to digital assets, but also for assistance to data creators in managing their own active data. Attention to current data management ensures not only that data can be preserved and reused by others, but also that creators can find their own data after its
initial use. Good management practices ensure that data can be discovered and validated if it is challenged or needs to be reexamined for any reason. Librarians who have worked with researchers on data transfers and documentation have found that recordkeeping practices within research teams are often idiosyncratic and inconsistent, at best. A 2012 survey at the University of Nottingham, for example, asked researchers in science, engineering, medicine and health sciences, and social sciences, “Do you document or record any metadata about your data?” Of the 366 researchers who responded, 24 percent indicated that they did assign metadata, 59 percent said no, and 17 percent did not know (Parsons, Grimshaw, & Williamson, 2013). Based on the results of the survey, the University of Nottingham Libraries are developing services to assist researchers with their needs for managing, publishing, and citing their research data.

Figure 1. Elements of digital repository services. Courtesy of Lars Meyer.

SCHOLARLY COMMUNICATIONS
Scholarly communications have moved beyond reliance on the published record in academic journals as the preferred way to share information among peers. Research results are now likely to be announced at professional conferences and in the news media, and data are shared through electronic communications among research teams that interact across
geographic boundaries and often across disciplines. In response, libraries have adapted and extended their services beyond preserving the published results of research to supporting the communications process throughout the data life cycle. Many research libraries are developing new services, including providing assistance with data management plans, helping with citations to published datasets—which are now beginning to appear in their own right in specialized online data journals—and managing repositories that preserve the datasets referenced in the citations.

Data journals provide quicker access to findings and underlying data in advance of published analyses that appear in “traditional” journals (which, however, are also likely to be issued in electronic form) and which serve the purpose of official documentation of research findings. See, for example, the Biodiversity Data Journal (motto: “Making your data count!”) at http://biodiversitydatajournal.com as one of these new types of e-journals. Many data journals, like the Biodiversity Data Journal, span a range of disciplines, so they have the advantage of presenting in one place datasets that bring together observational and experimental data, as well as analyses, from a variety of disciplines on a global spatial scale. Thus, data journals have particular value for the publication of interdisciplinary research.

In recognition of the growing significance of data publications, Nature Publishing Group (NPG) announced in April 2013 a new peer-reviewed, open-access publication, Scientific Data, to be launched in spring 2014. While the initial focus is on experimental datasets from the life, biomedical, and environmental sciences, there are plans to expand to other fields in the natural sciences. Scientific Data will introduce what it calls data descriptors, a combination of traditional publication content and structured information to be curated in-house, and which may be associated with articles from a broad range of journals. The actual data files will be stored in one or more public, community-recognized systems, or in the absence of a community-recognized system, in a more general repository such as Dryad (http://datadryad.org). An advisory panel including senior scientists, data repository representatives, biocurators, librarians, and funders will guide the policies, standards, and editorial scope of the new data journal (NPG, 2013). All of these professional groups bring specialized expertise to the scholarly communications process and are stakeholders in its successful evolution.
NEW FUNDING REQUIREMENTS FOR DATA MANAGEMENT PLANS

The NSF has had a long-standing policy requiring grant recipients to share their data with other investigators, but it had no policies for how this should be accomplished. Several significant reports published over the past decade have drawn attention to the need for a digital preservation infrastructure (Blue Ribbon Task Force, 2008 and 2010; National Science Board, 2005).

Awareness of the value of data took a leap forward in 2010, when the NSF announced that it would begin requiring data management plans with all grant applications beginning in the 2011 grant cycle. Research universities that depend heavily on NSF grant funding suddenly realized that the game had changed and that they would need to provide resources and assistance to researchers to enable them to compete successfully for grant funding. Other funding agencies in the United States and abroad soon began requiring data management plans also, so the need to act became critical.

Institutions have responded in different ways to the challenge of data management, based on their needs and circumstances. In many cases, libraries have played a critical role in the formulation of data management plans, bringing their knowledge of information standards and organizational skills to the process of setting up file structures, describing data in accordance with established metadata schemas and controlled vocabularies, and raising awareness of copyright, licenses, and other potential data rights issues. Many researchers have expressed willingness to share at least some of their data and have readily accepted assistance in managing their data for their own benefit as well as for sharing with others, as long as their concerns are met that data will not be shared inappropriately and that their work will not be slowed by cumbersome procedural requirements. A good data management plan will not only satisfy grant application requirements, but will also serve as a blueprint for instituting good practices for managing active data and facilitating long-term access. With training, much of this work can be carried out by the people on research teams who already have data management responsibilities, often graduate students and research assistants. It can be expected that many of the graduate students trained in good data management will go on to establish their own research teams and will promote good practices.
THE DATA LIFE CYCLE

Library and archival perspectives have come together in the past 10 years as the need to provide both good documentation and useful access to data and associated software tools has increased. It is now widely recognized that good management practices, in addition to data storage, are essential for successful long-term preservation and sharing. The skills needed to manage data effectively are now seen as spanning the library and archives professions; disciplinary expertise for understanding the specific data at issue is of course also required. This recognition of the need for collaboration across spans of expertise has led to the emergence of a new field known as digital (or data) curation, which can be succinctly defined as the active management of data over its full life cycle. The life cycle concept has helped focus attention on issues of data quality and documentation at the time of creation as critical to data-driven research, as well as for successful data preservation and sharing. The life cycle approach emphasizes the need for involvement of all stakeholders in the scholarly communications process, from those who create the data to those who manage and provide access to it over the long term.

Digital curation became a visible part of the digital knowledge environment in 2004 with the establishment of the Digital Curation Centre (DCC) in the UK (http://www.dcc.ac.uk). The DCC has provided leadership in promoting digital curation standards and best practices. A number of research universities in the United States—particularly research libraries—also have established digital (or data) curation centers and/or data services. These new organizations and service centers have played an important role in developing and supporting a community of data professionals, through such activities as the DCC’s International Digital Curation Conference and the International Journal of Digital Curation. In the United States, grant funding from the Institute of Museum and Library Services (IMLS), beginning in 2006, has supported the education of a cadre of digital curators by a number of graduate schools of library and information science; it also has provided funding for applied research in digital curation and information science. A study by the National Academy of Sciences Board on Research Data and Information on future career opportunities and educational requirements for digital curation, sponsored by IMLS, NSF, and the Alfred P. Sloan Foundation, is scheduled for release in late 2013.
Digital curators with backgrounds in librarianship, archival science, and related disciplines are contributing to the development of a new set of services that libraries and data service centers are now providing or contemplating.

Developments over the past decade have contributed to research data management in the United States and the roles that librarians and other information professionals are playing. This book provides a snapshot of the current state of the art, both for organizations that are considering such services and those that already provide them and wish to compare their own services with other initiatives. The contributors are all recognized experts in the field who have led the development of the first generation of data curatorship.

THE STRUCTURE OF THE VOLUME

The volume is organized to progress logically from considerations of the policy environment within which research data are created and managed, to the planning and implementation of services to support active data management and sharing, to the provision of archiving and repository services. These sections are followed by two contributions on evaluation planning (which, however, should be considered early in the life of the project or program, once decisions are made about the general goals and objectives and concurrently with the work plan). The last section includes case studies that serve as a link between the “what and why” issues discussed in earlier chapters and the challenge of “how” goals and objectives can be accomplished, presenting accounts of data services implemented at four research universities. The final contribution, by Clifford Lynch, puts the volume in context by considering where the field needs to go from here—not only the challenges that need to be solved in the next few years, but also the next set of challenges that will arise.

PART 1: UNDERSTANDING THE POLICY CONTEXT

This section provides a broad context for understanding how libraries in the United States have arrived at the current juncture between their historical roles and the changing environment of scholarly communications. It also describes innovative service models and strategies for influencing national and international policies to address legal and technological barriers to effective data management.
In “The Policy and Institutional Framework,” James Mullins provides an historical overview of the challenges faced by U.S. research libraries in the changing research environment of the past 15 years and how they have responded to evolving needs. He also provides a personal perspective on the Purdue University Libraries’ creation of its Distributed Data Curation Center and associated data services. These services are integrated with ongoing collaboration with faculty in order to meet their data management needs and to raise awareness of how librarians can support the university’s research and scholarly outputs.

MacKenzie Smith discusses the technology and policy context of data governance from a national and international perspective in “Data Governance: Where Technology and Policy Collide.” She describes the governance framework—the legal, policy, and regulatory environment—for research data and explains the ways in which it lags behind the established structure for traditional scholarly communications. She also discusses current efforts to resolve the legal, policy, and technical barriers to successful data management, and she offers suggestions for additional community-based tools and resources.

PART 2: PLANNING FOR DATA MANAGEMENT

This section discusses decisions that should be made at the beginning of the research process and the issues that should be considered in making them.

Jake Carlson, in “The Use of Life Cycle Models in Developing and Supporting Data Services,” compares the life cycle of data to life cycle models used in the life sciences, that is, identification of the stages that an organism goes through from birth to maturity, reproduction, and the renewal of the life cycle. He suggests that life cycle models provide a framework for understanding the similar stages of data and for identifying what services can be provided, to whom, and at what stage of the cycle. He cautions that gaps that may occur as data is transferred from one custodian to another require particular attention. However, these danger points in the life cycle present opportunities for services to mitigate loss of data or inadequate documentation.

Andrew Sallans and Sherry Lake, in “Data Management Assessment and Planning Tools,” discuss their work on the Data Management Planning (DMP) Tool, a community-developed resource maintained by the California Digital Library to help researchers establish a functional approach to man-
aging their research data while fulfilling grant application requirements; its successor, the DMPTool2; and DMVitals, developed by the University of Virginia Libraries. DMVitals combines a data interview with statements developed by the Australian National Data Service to describe best practices in data management. The tool enables researchers to score the "maturity level" of their current data management practices. Librarians can then provide recommendations for improving these practices and offer services to facilitate the process.

Bernard Reilly and Marie Waltz, in "Trustworthy Data Repositories: The Value and Benefits of Auditing and Certification," explain what it means for repositories to be considered trustworthy and how the Trustworthy Repositories Audit and Certification (TRAC) standard and checklist is used in making this determination. This chapter appears in Part 2 because the principles set forth in the TRAC document should be well understood by information professionals and considered early in the research planning process. While decisions about where to deposit research data at the end of their active life cycle may be made later, the TRAC criteria identify decisions that should be made before any data is created—such as assignment of unique identifiers and appropriate metadata—that are important for managing active research data and that also will facilitate deposit and sharing.

PART 3: MANAGING PROJECT DATA

This section presents aspects of project management around which information professionals can design services that build on their traditional areas of expertise to help researchers manage and share their data. These include considerations of copyright and licensing, provision of metadata services, and assistance with data citation. Libraries may consider offering such assistance either as stand-alone services or in combination with repository services.

"Copyright, Open Data, and the Availability-Usability Gap: Challenges, Opportunities, and Approaches for Libraries," by Melissa Levine, discusses copyright in terms of policy, administration, and business choices. She argues that librarians can help researchers achieve academic recognition and protect their data from inappropriate use through licensing (such as the Creative Commons-BY [Attribution] license) as an alternative to copyrighting their data. Levine proposes that assistance with decision making about
Rights in data is a logical addition to other data services that libraries may offer. Moreover, she cites the White House Office of Science and Technology Policy memorandum issued in February 2013, “Increasing Access to the Results of Federally Funded Scientific Research,” as a further incentive to researchers, librarians, and other stakeholders to continue and increase their collaborative efforts. The memo requires federal agencies that award more than $100 million for research and development annually to require data management plans with grant applications and provides for inclusion of appropriate costs to implement the plans. It further requires these agencies to “promote the deposit of data in publicly accessible databases, where appropriate and available” and to “develop approaches for identifying and providing appropriate attribution to scientific datasets that are made available under the plan” (Holdren, 2013, p. 5).

“Metadata Services,” by Jenn Riley, points out that metadata is a primary focus of data management plans. While funding agencies do not prescribe any particular metadata schemas, they expect researchers to adhere to the standards adopted by their own research communities and/or that best fit the data they are generating. She notes that metadata, like data, also has a life cycle. In addition to descriptive metadata that describes the content and provenance of the data, metadata will be added by machines or humans at later stages, including during preservation actions taken by repositories to enable access, citation, and reuse. Riley presents survey evidence showing that researchers are aware of the value of metadata yet are not knowledgeable about its proper application. She suggests that libraries can best provide effective metadata assistance by integrating services into the researchers’ workflow—thus increasing the benefit to the data creators—rather than waiting until the project's end, when researchers are unlikely to want to spend time documenting data they are no longer using.

“Data Citation: Principles and Practice,” by Jan Brase, Yvonne Socha, Sarah Callaghan, Christine Borgman, Paul Uhlir, and Bonnie Carroll, describes the development of and services provided by DataCite, an international consortium of libraries and research partners to encourage and support the preservation of research data as well as the citation of datasets to ensure their accessibility and to promote their use. The authors point out that data have been linked traditionally to the publications that are based
on them through tables, graphs, and images embedded in the publications. However, as datasets become larger, it often is no longer possible to publish the data as part of the publication. The datasets referenced in publications frequently are composite data objects with multiple constituent parts, as researchers typically generate many versions of datasets in the course of their research. The purpose of data citation, then, is to provide enough information to locate the referenced dataset as a single, unambiguous object; to serve as evidence for claims made about the data; to verify that the cited dataset is equivalent to the one used to make the claims; and to correctly attribute the dataset. The authors propose a list of 11 elements, ranging from author to a persistent URL from which the dataset is available, as the minimum required for data citation.

PART 4: ARCHIVING AND MANAGING RESEARCH DATA IN REPOSITORIES

This section focuses on the particular issues associated with data repositories. Libraries increasingly are involved as developers, service providers, and customers of such repositories, so they need to be knowledgeable about the range of repository models and services available. Contributors to this section describe a number of repository options, ranging from new roles for institutional repositories (IRs) in hosting active data, to new partnerships between disciplinary and institutional repositories as a means of improving archiving practices and making data more widely available, to emerging repository services offered by nonprofit organizations to accommodate a wide variety of content.

In “Assimilating Digital Repositories into the Active Research Process,” Tyler Walters makes the case for IRs as infrastructure to support large research projects. These projects, often involving international teams of researchers from many disciplines, are now typical and require a networked research environment. Walters observes that repositories are being integrated with the communication tools of virtual communities and that social media tools and community networking capabilities are overlaying repositories to link data, people, and web-based resources. He argues that in order to benefit researchers, digital repositories should play a larger role in supporting active research in addition to archiving data.
In “Partnering to Curate and Archive Social Science Data,” Jared Lyle, George Alter, and Ann Green discuss the exponential increase in the volume of social science research data in recent years and the potential loss of much of this data through lack of proper archiving. The authors provide evidence that the vast majority of social science research data are currently shared only informally or never shared beyond the original research team. They recognize the valuable role that IRs are playing in capturing inactive research data and suggest that disciplinary repositories such as the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan can improve archiving practices and data sharing by partnering with IRs. They report on the results of an IMLS grant to the ICPSR to investigate the possibilities for partnerships between the ICPSR and IRs, which typically serve as general repositories for a university’s scholarly outputs. The project found that many IR managers were receptive to suggestions for improving documentation of social science data and that the ICPSR could successfully obtain relevant datasets from IRs, making them more easily discoverable by social science researchers. The chapter concludes with recommendations for improvements in archiving practices that are relevant not only for IRs, but for all those involved in managing research data, especially information professionals.

In “Managing and Archiving Research Data: Local Repository and Cloud-based Practices,” Michele Kimpton and Carol Minton Morris discuss practical considerations for making decisions about what kinds of data to preserve in repositories, for how long, and in what kinds of repositories. They also provide insight into commercial cloud-based storage practices, which are often opaque to users. The first part of the chapter presents an analysis of four recent interviews with research library professionals who use either the DSpace and/or Fedora repository software. The interviews were conducted to better understand the common issues and solutions for preserving and using research data in local repositories. The second part discusses the challenges and benefits of using remote cloud storage for managing and archiving data, including considerations of data security, cost, and monitoring. The authors describe the DuraCloud service provided by DuraSpace as a resource designed to overcome the opacity of commercial cloud-based services.
“Chronopolis Repository Services,” by David Minor, Brian Schottlaender, and Ardys Kozbial, describes the repository services provided by the Chronopolis digital preservation network, created and managed by the San Diego Supercomputer Center (SDSC) and the University of California-San Diego Library in collaboration with the National Center for Atmospheric Research in Colorado and the University of Maryland Institute for Advanced Computer Studies. The network takes advantage of its distributed geographical locations to ensure that at least three copies of all datasets deposited with Chronopolis are maintained, one at each of the partner nodes. Data is managed with the iRODS (integrated Rule-Oriented Data System) middleware software developed at the SDSC and is continually monitored through “curatorial audits.” Chronopolis is a “dark archive,” meaning that it provides no public interface and only makes data available back to the owners; however, it has developed model practices for data packaging and sharing through its ingest and dissemination processes. It promises to become a useful component of digital preservation for a wide variety of content.

PART 5: MEASURING SUCCESS

The contributions here emphasize the need to begin planning for evaluation at the beginning of a new project or program. However, these chapters follow the earlier sections because decisions about what services to provide must be made before evaluation planning can begin. The authors in this section consider evaluation from two perspectives. The first provides an in-depth analysis of the steps involved in developing and implementing an evaluation plan for a large, complex, data-focused project with several goals and many stakeholders. The second takes a high-level view of evaluation as a means of assessing the return on investment of public funds to meet national or international goals.

In “Evaluating a Complex Project: DataOne,” Suzie Allard describes the planning and evaluation of DataONE, a multimillion-dollar project funded by NSF’s DataNet program. The goal of DataONE is to develop infrastructure, tools, and a community network in support of interdisciplinary, international, data-intensive research spanning the biological, ecological, and environmental sciences. While this project is large and complex, requiring particular care in planning for project evaluation, many of the evaluation
components will have relevance for any project that intends to measure outcomes, and particularly for those involving management of research data. Allard emphasizes the importance of developing an evaluation plan in the early stages of the project in order to ensure that relevant data are collected at appropriate times. She also explains how the data life cycle model helped to structure the DataONE evaluation plan, since the ultimate project goal is to improve data management. This framework helped to identify the tools and resources needed at each stage of the life cycle. The evaluation team then developed plans for evaluating existing tools to assess potential improvements and for identifying needs for new tools and services that could be addressed in the work plan. Allard concludes with recommendations for the organizational design and management of the evaluation process.

In “What to Measure: Toward Metrics for Research Data Management,” Angus Whyte, Laura Molloy, Neil Beagrie, and John Houghton discuss evaluation metrics from a high conceptual level, asking program and evaluation planners to think carefully about what they are trying to achieve and what metrics they can realistically use to measure results. The authors address the evaluation of research data management at two levels: the services and infrastructure support provided by individual research institutions, and the economic impacts of national or international repositories and data centers. Using cases from the United Kingdom and Australia, they consider methods such as cost-benefit analysis, benchmarking, risk management, contingent valuation, and traditional social science methods including interviews, surveys, and focus groups. However, they observe that the starting point should always be “what can and should be measured.” They remind readers that the goal is to identify improvements that have been achieved or are needed to align services with national or international data policies and practices. The authors note that, although data preservation is now perceived as a public good, the public benefit has not yet been proved, presenting particular challenges for evaluation.

PART SIX: BRINGING IT ALL TOGETHER: CASE STUDIES

This section presents case studies that describe how all of the policy, planning, and implementation considerations have come together in new services at four research universities.
Cornell University

In “An Institutional Perspective on Data Curation Services: A View from Cornell University,” Gail Steinhart notes the early interest that Cornell took in research data beginning in the 1980s and describes the planning and implementation of new library infrastructure and data services over the past two decades. She discusses important lessons learned from this wealth of experience and makes recommendations for structuring the planning and ongoing monitoring processes that are essential to successful data services.

Purdue University

In “Purdue University Research Repository: Collaborations in Data Management,” Scott Brandt extends the observations made by Purdue’s Dean of Libraries James Mullins. Brandt provides insight into how Purdue librarians acted within the policy and institutional framework described by Mullins. He emphasizes the value of collaboration with researchers throughout the data life cycle for librarians who are continuously working to improve data services.

Rice University

Geneva Henry’s case study, “Data Curation for the Humanities: Perspectives from Rice University,” is the only chapter that focuses on humanities research data, an important but often overlooked area in research data management. Scientific data have received the most attention in the development of data services because this area has led the transition to digital research and receives the bulk of research funds. However, the digitization of historical data and the development of data mining and other techniques for analyzing humanities data are now enabling researchers in the humanities to make innovative use of digital tools and to ask research questions that were not possible before.

University of Oregon

Brian Westra, in “Developing Data Management Services for Researchers at the University of Oregon,” describes the development of data services, particularly for the sciences and social sciences, at his large, state-supported research university. He provides a detailed description of a needs assessment conducted in 2009–2010 that identified common problems—including
lack of file organization, insufficient storage and backup procedures, and inadequate metadata or other documentation—that hindered investigators’ ability to find and retrieve their own data. These conditions aligned with gaps in infrastructure, tools, and services that the library then undertook to fill. Westra discusses the use of small pilot studies early in the data life cycle to explore collaborations in the development of data management infrastructure.

CLOSING REFLECTIONS: LOOKING AHEAD
Finally, Clifford Lynch contributes a thoughtful reflection on progress to date in preserving and managing research data. He identifies priorities over the next few years for services to assist researchers with data management, centering on: development of credible data management plans, documentation of datasets to be shared and preserved, and appropriate platforms for data sharing and bit preservation. He further considers what these developments and the current environment suggest for the next set of challenges that will inevitably arise.

CONCLUSION
Institutions have responded differently to the new data environment depending on their own circumstances and needs. Therefore, a variety of approaches are presented in this volume, and information professionals looking to develop strategies for their own institutions will find many examples to choose from and adapt to their own needs. The data infrastructure is still emerging, but there are many more tools and services available now than there were ten or even five years ago. Libraries have played a critical role in developing and managing this infrastructure and are likely to become even more involved as research becomes ever more dependent on digital data. Collectively, the state of the art as described here demonstrates the resilience of libraries and information professionals in responding to the changing needs of their communities.

REFERENCES


