## Contents

**Introduction to Research Data Management**  
*Joyce M. Ray*

**PART 1: UNDERSTANDING THE POLICY CONTEXT**

1. The Policy and Institutional Framework  
   *James L. Mullins*  
   25

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

**PART 2: PLANNING FOR DATA MANAGEMENT**

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

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

5. Trustworthy Data Repositories: The Value and Benefits of Auditing and Certification  
   *Bernard F. Reilly, Jr., and Marie E. Waltz*  
   109
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's Next for Data? A Vision for the Future 261

PART 6: METADATA SERVICES

15 An Introduction to Metadata: A World of Description, Categorization, and Accessibility 275
Jover 

16 Pure Collaboration: Building Data Commons 287
Debra Katz and Mary K. Lambert

17 Data Curation for Professional Librarians 303
George Alter

18 Developing Metadata for Public Libraries 321
Bonnie Carroll

CLOSING

19 The Future of Archiving and Managing Research Data: Challenges and Opportunities 337
About the Authors
Index
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

About the Contributors

Index
INTRODUCTION

Interest in and support for cyberinfrastructure (CI) development and data-driven research has grown significantly in the past decade or so. Well-curated and accessible data are the feedstock for CI and data-driven research. As custodians of the scholarly record, research libraries might be presumed to have an important role as curators of research data. Libraries are not entirely new to supporting research data in other capacities, with roles such as geographic information systems and social science data librarians being fairly common.

In spite of this seemingly natural fit, introducing and operating data services is a complicated undertaking. Publishing processes, models, and infrastructure are well established and well understood, despite some tension and change in the transition from print to digital formats as well as experiments in open access and alternative revenue models. Researchers do their work, write and submit papers, participate in the peer review process, and libraries acquire and make available the resulting literature. Research data present would-be custodians with a more complicated set of challenges. The definition of data curation used by the Graduate School of Library and Information Science at the University of Illinois (n.d.) nicely captures some of the complexity:

Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarly and educational activities across the sciences, social sciences, and
the humanities. Data curation activities enable data discovery and retrieval, maintain data quality, add value, and provide for re-use over time.

This definition encompasses and implies continuing management throughout the data life cycle, the need for discipline-specific expertise to add value and ensure quality, support for discovery and access, and preservation of content and access to it over the long term.

The problems are not merely technical; research data curation involves the interplay between multiple, distributed technological and sociocultural systems. From the perspective of individual researchers, research funders are moving steadily toward requiring the sharing of research data, yet researchers are often reluctant to share due to concerns over intellectual property, attribution, improper reuse, and lack of time, resources, and know-how to get the job done. At the institutional level, there exists considerable uncertainty over custodial roles—who should do what, how, for how long, and who should pay. There is particular tension as to the roles and responsibilities of disciplines and institutions such as universities and their libraries. At the broadest scale, intelligent curation services should support emerging CI, but the interplay of local versus global as well as organic and more organized CI developments “challenge simple notions of infrastructure building as a planned, orderly, and mechanical act” (Edwards, Jackson, Bowker, & Knobel, 2007) and make moving targets of optimal CI development and data curation strategies.

Addressing this suite of challenges is a tall order. Planning and decision making for research libraries wishing to engage in this area can seem daunting, particularly as opportunities and needs arise while or even before planning activities are completed. With the hope that others might learn from Cornell University Library’s (CUL) experiences, I present an overview of planning activities, data services and activities (past, current, and planned), practical lessons learned, and a few ongoing challenges.

PRECURSORS TO RESEARCH DATA CURATION SERVICES
Cornell’s interest in support for research and for digital data dates at least to the early 1980s, with the start-up of the data archive at the Cornell Institute for Social and Economic Research (CISER), Albert R. Mann Library’s hiring of its first nonbibliographic computerized data files librarian, and
Enables data discovery, enables data discovery, and delivers continuing management of discipline-specific expertise to discovery and access, and preservation for the long term.

Sensible; research data curation involves distributed technological and socio-cultural, individual researchers, researching the sharing of research data, due to concerns over intellectual property and lack of time, resources, and institutional level, there exists concern—who should do what, how, for particular tension as to the roles institutions such as universities and libraries, and intelligent curation services should take local versus global as well as other “challenge simple notions of infrastructures” (Edwards, 1992). I make moving targets of optimal CI a tall order. Planning and decision making to engage in this area can seem overwhelming. Needs arise while or even before the hope that others might learn from the experiences, I present an overview and activities (past, current, and upcoming) ongoing challenges.

CURATION SERVICES

The development at Mann Library of an experimental system called the Interactive Numeric Files Retrieval System (INFeRS) to provide access to numeric files in the agricultural and life sciences (Figure 1). The 1990s saw the hiring of CUL’s first metadata librarian at Mann Library, the launch of the USDA Economics, Statistics, and Market Information System (USDA-ESMIS, http://usda.mannlib.cornell.edu/), and the development of the Cornell University Geospatial Information Repository (CUGIR, http://cugir.mannlib.cornell.edu/, one of two geospatial data clearinghouses for New York State). The USDA-ESMIS began with an informal conversation in 1992 as to whether the Mann Library could host files from one of its agencies, the Economic Research Service. Twenty years later, Mann provides free and timely online access to data and reports for five USDA agencies. CUGIR was initially created for the purpose of distributing U.S. Census TIGER/Line data online, and funded with a National Spatial Data Infrastructure Cooperative Agreements Program grant. Since its inception, the CUGIR collection has grown to include over 7,500 datasets from 17 partners, with an average of more than 85,000 downloads per year. More recently, the CUGIR team has been encouraging Cornell researchers to use the repository to publish their GIS datasets.

In the 2000s, CUL assumed responsibility for the physics preprint repository, arXiv.org, and established a Metadata Services Unit within Library Technical Services. Around the same time, Mann Library began to develop VIVO, a research and expertise discovery system based on semantic web technologies (Porciello, Devare, & Corson-Rikert, 2008). VIVO development continued at a multi-institutional scale from 2010–2012 with a grant from the National Institutes of Health (NIH); more recently, VIVO has joined the DuraSpace Organization Incubator Initiative, a move that will ensure community support and development.

Cornell’s institutional repository, eCommons, was launched by former Dean of the Faculty Robert Cooke in 2002, with funds from the Atlantic Philanthropies. Cooke’s intention was to have the project start as a faculty-centric (rather than library-centric) one, and to use the DSpace-based repository to support the open access scholarly publishing prototype, the Internet-First University Press. At the conclusion of the funded project, CUL assumed responsibility for the ongoing operation of the repository and has managed it ever since. Cooke correctly forecast the use of the repository...
Figure 1. Timeline of data services and planning activities.

- 2013: First digital collections added to CULAR
- 2012: ARLESI e-research agenda completed
- 2011: First NSF data management plan consultation completed
- 2010: Research Data Management Service Group (RDM) proposed
- 2009: Data Executive Group formed
- 2008: Data Discussion Group started
- 2007: NERI grant for VIVO awarded
- 2006: Data Working Group white paper completed
- 2005: DISCOVER research services formed
- 2004: Second NSF-CI grant awarded (Datarat)
- 2003: First NSF-CI grant awarded (SGER to Mann Library)
- 2002: VIVO development begun
- 2001: Metadata Services established
- 2001: Commons started
- 2000: arAv moved to CUL
- 1998: CUGIR established
- 1997: First metadata librarian hired
- 1993: USDA-ESMIS established
- 1989: INFERN project begun
- 1984: First Non-bibliographic Computerized Files Librarian hired
- 1982: CISER data archive established

For additional data services and planning activities, see the timeline provided.
for content such as research data, computer code, and other supplementary materials not typically supported at that time by conventional publishers. As of this writing, eCommons contains 86 items classified as datasets. CUL’s continued development of an infrastructure for digital preservation led to the CUL Archival Repository (CULAR), a preservation repository intended to support the long-term preservation of CUL’s digital collections, including eCommons. The first digital collections to be added to CULAR were processed in February 2013.

While these systems and activities were not focused primarily on the collection, management, and distribution of research data, some have been used in that capacity. Developing these services has helped establish CUL’s capacity for supporting data curation with infrastructure and with expertise, positioning the library to work effectively in this arena.

PLANNING FOR DATA SERVICES AT CORNELL UNIVERSITY LIBRARY

Early on, CUL developed relatively specialized data infrastructure and services as opportunities arose. It was not until funders’ interest in supporting CI development and data curation became clearer that CUL undertook its first systematic planning initiatives, and since the mid-2000s, planning activities and development of infrastructure and services have proceeded more or less in tandem. Figure 1 provides a timeline view of planning activities and service developments discussed in this chapter.


Established in 2006, the Data Working Group was charged with reviewing the current research data curation landscape and with recommending strategic opportunities for CUL. The group included subject area and collection development librarians, Library Technical Services and Information Technology staff, policy experts, and staff from the Weill Cornell Medical College. They reviewed several of the key reports on CI and data curation, and undertook environmental scans of activities within CUL, at Cornell, and beyond. The group also reviewed the current practices and trends of selected disciplines and at other research libraries, educational opportunities for professionals, and selected critical issues related to data curation and sharing.
The Data Working Group made the following recommendations in its final report (Steinhart et al., 2008):

1. **Seek out and cultivate partnerships with other organizations.** This was perhaps the group's most important recommendation, and they suggested that CUL partner with campus units such as the Center for Advanced Computing and Cornell's central Information Technology division.

2. **Provide services to Cornell researchers in several areas.** At the time, the National Science Foundation's (NSF) intention to require data management plans was becoming clear but had not yet gone into effect. The Data Working Group recommended trying to educate Cornell researchers on emerging requirements and helping researchers to meet them. Other opportunities for supporting researchers included providing information on best practices in managing data, referring researchers to appropriate sources of information and expertise, and participating in the formulation of institutional policies on research data retention.

3. **Assess local needs and develop local infrastructure and related policies.** The Data Working Group recommended that CUL work with other stakeholders to identify and articulate researchers' needs for data management infrastructure and participate in its development. The group further recommended the development of selection and appraisal guidelines, an investigation of the role of the library in curating data related to publications, and continued work in the curation of small-scale datasets.

4. **Cultivate a workforce capable of addressing the new challenges posed by data curation and CI development.** This recommendation sought to encourage the identification and development of skills needed and to extend CUL's activities in this area.

5. **Form a Data Curation Executive Group and reorganize the Data Working Group.** The working group recommended the formation of a permanent committee to guide CUL's data curation and CI activities, and a separate group to foster a professional community and development for librarians engaging in this work.

CUL successfully implemented each of these recommendations to some degree, collaborating with other units to explore researchers' needs and to launch new services, experimenting with the development of Datastar (described later in this chapter), developing basic policies for depos-
owing recommendations in its

with other organizations. This was
commendation, and they suggest-
such as the Center for Advanced
ation Technology division.
areas. At the time, the
ntention to require data manage-
d not yet gone into effect. The Data
educate Cornell researchers on
searchers to meet them. Other op-
ents included providing information
ferring researchers to appropriate
 in curating data related to publica-
iting the new challenges posed by
This recommendation sought to en-
ment of skills needed and to extend

and reorganize the Data Work-
ommended the formation of a per-
ata curation and CI activities, and a
nal community and development for

ach of these recommendations to
its to explore researchers’ needs
ng with the development of Data-
developing basic policies for depos-
iting data in the institutional repository, hiring additional librarians with
data responsibilities in their job descriptions, and forming the Data Execu-
tive Group and the Data Discussion Group.

Association of Research Libraries E-Science Institute
(ARLESI, 2010–2011)

With the fall 2010 announcement that the Association of Research Librar-
ies (ARL) and the Digital Library Federation (DLF) would be collaborat-
go the first ever E-Science Institute, a guided series of planning exer-
ces resulting in the development of an agenda for e-research by each par-
ticipating institution, the time seemed right for CUL to revisit and update
its plans with respect to data curation and e-research support in general. A
team made up of selected members of the Data Executive Group developed
the agenda, which included 21 possible initiatives. CUL’s executive leader-
ship group identified several as medium to high priority:

1. Participate in development of institutional data policies, including Cor-
nell’s draft data retention policy. The Office of the Vice Provost for Re-
search was the original sponsor of the draft data retention policy for Cornell.
The original Data Working Group had an opportunity to comment on the
original draft policy, but due to changes in leadership and priorities in the
Vice Provost’s office, the policy has been tabled for the foreseeable future.
2. Position the Research Data Management Service Group (RDMSG, dis-
cussed in greater detail later in this chapter) as a resource throughout
the life cycle of a research project. This objective encompassed working
to ensure the continued success of the RDMSG, including ensuring aware-
ness of the RDMSG among researchers, and reviewing information CUL
and its partners had already collected to ascertain unmet needs for re-
search support.
3. Define CUL’s role in data preservation in relation to Cornell partners.
The library is generally seen as the natural archive for datasets for the long
term, and yet the library lacks the expertise and infrastructure to effective-
ly preserve some of the large and complex types of data being generated by
Cornell researchers. This objective prompts CUL to define and articulate
its capabilities and role to its partners, so that additional capacity can be
developed if needed.
4. **Enhance CUL's metadata services for data.** This activity includes extending VIVO to accommodate basic dataset description (the Datastar project), drafting best practices guidelines for depositing datasets in eCommons, and participating in specific research data projects as the opportunity arises.

5. **Inventory current skills, responsibilities, and positions as they relate to e-research.** Identify desired skills that are not present or could be further developed, in order to sustain or improve that balance as positions are revised or created. This was a priority of the original Data Working Group, and the ARLESI team viewed this as an ongoing need. The library as a whole, as part of its strategic planning process, charged a team with defining a protocol for developing a "competency profile" for all groups of staff positions.

6. **Identify and provide professional development opportunities for existing staff to improve their skills related to e-research support.** CUL staff are encouraged to take advantage of professional development opportunities as they arise. For example, CUL recently sent its life sciences librarian to a workshop on supporting translational and team-based science.

7. **As CUL's staff expertise grows and as opportunities arise, partner with researchers to support discipline-specific support for e-research.** CUL continues to do this on an ongoing basis, participating in the development of proposals to such agencies as the U.S. Environmental Protection Agency, the NSF, the National Endowment for the Humanities, and the New York State Energy Research and Development Authority.

8. **Determine which CUL services are free or fee-based, in order to develop costed and sustainable models and fee structures.** As libraries are asked to provide new services and infrastructure—without an increase in funding or ceasing to provide existing services—it will be necessary to seek sources of revenue in order to provide data curation and e-research services.

Several other recommendations in the strategic agenda were already planned or underway, and so were not included in the priority list above. Among these, the RDMSG was identified as an important component of the library’s strategy for supporting researchers. More specifically, we anticipate a need for data management planning outreach and training specific to researchers working with the NIH. Currently, data management plans are only required of NIH projects receiving more than $500,000 of funding in a single year; when the requirement becomes more broadly applicable, the RDMSG will
initiate outreach and training activities similar to those undertaken for NSF grant writers. Guidance and training for researchers (including graduate students) in best practices for data management planning and for data management was also identified as important. Cornell is a partner in Purdue University’s Data Information Literacy project (http://wiki.lib.purdue.edu/display/stc), which aims to develop data management curricula for graduate students across several disciplines. Cornell’s curriculum is offered as a one-credit course for graduate students in natural resources and related fields.

Finally, the agenda included a recommendation to examine the options for supporting data deposit in conjunction with articles uploaded to arXiv.org, which is managed by CUL. A small team is currently examining the results of a pilot project that allowed researchers to upload datasets associated with publications in arXiv.org to the DataNet-funded Data Conservancy (http://dataconservancy.org/), as well as the use of arXiv’s capability to accept ancillary files as a means for distributing datasets.

Overall, we found the E-Science Institute to be a very productive way to revamp our data and e-research plans. Some team members were somewhat surprised that the exercise did not result in a recommendation to build a new general, campus-wide curation infrastructure, nor did it surface clear needs for other large-scale or overarching activities. Instead, most recommendations focused on matters of policy, staffing, continuing and enhancing existing services, seeking out partnerships, and developing business models. More or different interviews might have surfaced additional opportunities and concerns, but our results might also be what one would expect for institutions already actively providing data services, as well as a reflection of the complex and distributed nature of research and institutions.

**Concurrent Campus-Wide Planning Initiatives (2006–2010)**

There were two campus-wide efforts in roughly the same time frame as CUL’s that were aimed at collaboration and planning for the support of CI and data-driven research: a CI task force formed in 2006, and the Data Intensive Science Organization for Virtual Exploration and Research (DISCOVER) Research Service Group, proposed in 2008. The goal of the CI task force was to assess and make the case for upgrades to Cornell high-performance computing (HPC) facilities in order to make Cornell more competitive for CI grants, and the objective of seeking out CI grants helped
shape the 2007 reorganization of Cornell’s HPC center. Several library staff participated in the meetings of the CI task force, and while the library played only a minor role in the deliberations and activities of the group, its participation served to make the library’s interest in data curation known within Cornell’s HPC community.

The DISCOVER Research Service Group was intended to further investigate the needs of Cornell researchers with respect to HPC and data curation, and to explore solutions that might benefit multiple communities. The group consisted primarily of staff and faculty from the Center for Advanced Computing, academic departments, and the library. The Office of the Vice Provost for Research provided two years of funding to support two DISCOVER staff, with the intention that a successful service group would become self-sustaining by obtaining grants or developing fee-based services. DISCOVER staff conducted a dozen in-depth interviews with Cornell researchers and synthesized their concerns into a set of recommendations addressing storage, network, training, and data stewardship, all of which have or are being addressed by the DISCOVER partners. Significantly, it was the set of collaborative relationships established within the DISCOVER Research Service Group that eventually led to the formation of the RDMSG.

IMPLEMENTING DATA SERVICES AT CORNELL UNIVERSITY LIBRARY
Data services at CUL have evolved over time in tandem with changes in research practices and technology, opportunities to pursue specific projects, and the emergence of policies related to research data. Some services and infrastructure have been developed opportunistically, while others have been developed more strategically and as result of the planning exercises described previously.

Early Research Data Curation Efforts at CUL
The mid-2000s saw two separate collaborations with researchers on domain-specific support for curating research materials. In 2004, Mann Library and a faculty member of the College of Human Ecology received a small grant for exploratory research from the NSF to fund work investigating possible collaboration between research scientists and the library. The work focused on two research groups: the Cornell Language Acquisition Laboratory (CLAL) and the Upper Susquehanna River Basin
An Institutional Perspective on Data Curation Services

A Staging Repository for Research Data (Datastar)

Datastar was originally envisioned as a platform for user-controlled sharing of research data, the creation of basic metadata (and optionally, more detailed domain-specific metadata, for selected schemas), and a means for publishing datasets to supported external repositories (Khan et al., 2011; Steinhart, 2010). Our goal was to position the library in the research process without requiring interaction with the library if researchers were self-sufficient in using the system. Datastar was built using the Vitro software, the semantic web application and ontology editor that also supports the VIVO application. Datastar required significant additional development to integrate domain-specific metadata schemas, manage complex permissions rules, integrate a Fedora repository, and support the publication of datasets to external repositories. A prototype system with support for CLAL linguistic metadata, selected portions of the Ecological Metadata Language (EML) schema, and metadata for the Data Conservancy and for CUL’s eCommons was developed successfully. Proof-of-concept publication workflows to move data and metadata from Datastar to the Data Conservancy, a test instance of eCommons, and a local installation of Metacat (the data and metadata database that supports the Knowledge Network for Biocomplexity...
(http://knb.ecoinformatics.org/), an ecological data repository) were also developed and tested. While these accomplishments were encouraging, the development of usable interfaces to support creation and editing of metadata for multiple schemas was far too labor-intensive to be scalable and sustainable, and publication workflows could not be developed for some end-point repositories of interest.

With funders’ statements that proof of compliance with data-sharing policies would be expected in interim and final reports, the project team turned their attention to supporting basic data registry functions in a simpler, scaled-back version of Datastar. In collaboration with Washington University in St. Louis and supported with funding from the Institute of Museum and Library Services, the team completed a set of Data Curation Profiles (http://datacurationprofiles.org/, which are investigations based on the use of an interview instrument to gather detailed information about the full life cycle of a dataset) to inform the development of Datastar. Using the information gathered in the profiles, and inspired by similar system development at the University of Melbourne in support of the Australian National Data Service, the team is in the process of developing the Datastar software for open-source distribution.

A Campus-Wide Research Data Services Partnership: The Research Data Management Service Group

The RDMSG (http://data.research.cornell.edu) is the most coordinated and current effort to provide research data services to Cornell researchers. With the May 2010 announcement that the NSF would soon require a data management plan (DMP) with every grant proposal, individuals from multiple centers and departments, including the library, began to design a new service group to assist researchers in meeting the NSF requirement. The potential impact for Cornell was significant, with researchers submitting more than 400 proposals to the NSF per year. The NSF had not yet provided specific guidance at the time the group began to plan, but the Interagency Working Group on Digital Data, a group of two dozen federal agencies working to maximize the return on investment from their funded research, had issued a 2009 report that included specific recommendations as to what funding agency DMP policies should include. Using that report as a basis, the Cornell group was able to infer the types of support research-
ological data repository) were also accomplishments were encouraging, to support creation and editing of too labor-intensive to be scalable workflows could not be developed for.

Of compliance with data-sharing and final reports, the project team basic data registry functions in a simple.

In collaboration with Washington with funding from the Institute of Science, which are investigations based to gather detailed information about on the development of Datastar. Us-


data services to Cornell researchers that the NSF would soon require a yearly grant proposal, individuals from within the library, began to design a plan in meeting the NSF requirement. Significant, with researchers submitting NSF per year. The NSF had not yet the group began to plan, but the In-

Data, a group of two dozen federal agencies on investment from their funded included specific recommendations should include. Using that report infer the types of support research-

ers would likely need to know about to complete a competitive DMP. The group identified the Cornell units with services and expertise to support data management, identified some gaps in services, and proposed a model for a collaborative, virtual organization to support researchers at Cornell (Block et al., 2010).

Jointly sponsored by Senior Vice Provost for Research Robert Buhrman and University Librarian Anne Kenney, the RDMSG is managed by a coordinator and a management council, and it has a faculty advisory board (Figure 2). Management council members represent the main service units and allocate staff to serve as members of implementation teams, some of which are ongoing and some of which are convened to meet specific needs. Ongoing teams include the RDMSG consultants, the web and documentation team, and the outreach and training team. The consultant group includes individuals with disciplinary expertise, technical expertise, and policy expertise, and members come from the library, Cornell’s central information technologies unit, Weill Cornell Medical College, CISER, and academic departments.

**Figure 2.** Organization of the Research Data Management Service Group (RDMSG).
Early on, the RDMSG conducted a survey (Steinhart, Chen, Arguillas, Dietrich, & Kramer, 2012) of NSF-funded principal investigators, the results of which showed a need for advisory services for data management planning. Some of the key findings of the survey were that researchers were generally uncertain as to whether their data or metadata conform to standards in their discipline (or even whether such standards exist), and that many do not create metadata to document their datasets. The majority expressed an interest in data management planning assistance. There was moderate interest in guidance on issues of intellectual property and copyright, and (some comments to the contrary notwithstanding) a general willingness to share at least some datasets. Most respondents indicated that they would share relatively small amounts of data (77 percent said they would share 100GB or less), but the diversity of reported types and formats of data was impressive, with respondents listing 77 unique file extensions. These results suggest that service providers can expect to deal with very diverse but generally small datasets, and that researchers will need some assistance in adopting best practices for data management.

The RDMSG has been successful in reaching out to researchers, assisting them with DMPs, and connecting them to data management services on campus. In the first two years, the outreach and training team conducted 30 information sessions on general data management planning or topics for more specific audiences, such as individual academic departments or graduate students, with a total attendance of over 500 individuals. Feedback from session participants was overwhelmingly positive. The RDMSG consultant group also handled more than 100 individual consultations, most of which involved advising on DMPs. Into its third year, outreach efforts will shift from general information sessions to more specialized sessions for targeted audiences, as well as the development of content to support specific curation tasks, such as best practices for depositing research data in the institutional repository.

LOOKING FORWARD: OPPORTUNITIES AND CHALLENGES

CUL's work in data curation has evolved organically over time to consist of three main components: consultation services, core infrastructure, and specialized projects. Specialized projects with individual researchers and research groups who are open to some degree of experimentation have provided an extremely
An Institutional Perspective on Data Curation Services

In a survey (Steinhart, Chen, Arguillas, and principal investigators, the respondents for data management services identified that the survey were that researchers are not data or metadata conform to neither such standards exist), and in the metadata to document their datasets. The major issues of intellectual property and contrary notwithstanding) are datasets. Most respondents indicated that files of data (77 percent said they are not over what a listing 77 unique file extensions, whereas can expect to deal with a lot that researchers will need some data management.

Reaching out to researchers, assistance to data management services on teaching and training team conducted 30 management planning or topics for each academic department or grad-faculty and over 500 individuals. Feedback was overwhelmingly positive. The RDMSG conducted 200 individual consultations, most of its third year, outreach efforts to more specialized sessions of data management planning assistance. There is no particular management planning assistance. These include continuing to improve local services and infrastructure, providing organizational and staff support for data services as they evolve, refining the scope of data services to be provided, and dealing with uncertainty in a rapidly changing research, policy, and technological landscape.

Improving Existing Infrastructure

As noted above, early experience with data services has given us some insight as to how we might improve existing infrastructure. To that end, we recently revisited information gathered in earlier efforts to understand researchers' needs for local data management services and infrastructure. In addition to the RDMSG survey described earlier, these sources included the collection of Data Curation Profiles conducted to support the redevelopment of Datastar, and anecdotal information gathered through the RDMSG consultants' work with researchers. We found that support for a number of repository functions would likely be very helpful to researchers, and some are either already supported or are reasonably attainable with current or planned systems. These include:

- robust linking between datasets and the publications based upon them,
- version management for datasets,
- support for citations, including cut-and-paste citations and/or citation export function for data users, prepublication citation information allowing data owners to cite datasets in manuscripts ahead of the actual availability of data, and the ability to cite a specific version of a dataset,
- ability for data owners to control access to data for groups or individuals,
- policies and technical infrastructure to allow for updates to datasets that change or grow over time,
- record scheduling functions to allow for planned deaccessioning of content (depending on library and university policy developments in this area).
We also know from this exercise that we can expect to be challenged to curate highly diverse materials including less common and proprietary file formats, that quality metadata may be lacking or difficult to come by, and that some data will have privacy and confidentiality issues. Perhaps most significantly, if we decide it is necessary to recover some or all of the costs associated with data curation, we know that researchers are concerned with whether and how those costs will be recovered, particularly for services needed beyond the end of a research grant.

Providing Organizational and Staff Support for Data Services

Deciding where data “fits” in the library is not entirely straightforward. Data services inevitably engage expertise and resources in multiple functional areas, such as public services, information technology, and scholarly communication. The first action many libraries take is to hire a data librarian, often within a public services department, to develop data curation services, and Cornell was no exception. Since that initial appointment, Cornell has incorporated research data services into the job descriptions of additional subject area librarians working in public services, and the university also has refactored a metadata librarian position so that its focus is on science data and metadata. These are promising developments, but elsewhere it remains common for librarians working with research data to find themselves doing so in relative isolation, encountering difficulty in marshaling the resources needed to launch new services and develop infrastructure. When libraries are able to devote resources to develop infrastructure locally, the work is frequently experimental and funded by grants. Some of these projects are more successful than others, but even for successful projects, transitioning from project funding and support to a production-scale service and program can be quite difficult, and many efforts do not progress beyond the proof-of-concept stage.

In addition to staffing to deliver data services and developing infrastructure, it is imperative that data services be a part of a senior administrator’s portfolio, in order to ensure meaningful support to set direction and allocate resources for service development. Because research data services are still perceived as new and specialized, they often aren’t fully integrated into libraries’ organizational structures. At Cornell, primary responsibility for data curation services rests with the Chief Technology Strategist,
we can expect to be challenged to
less common and proprietary file
backing or difficult to come by, and
confidentiality issues. Perhaps most
of all, to recover some or all of the costs
that researchers are concerned with
covered, particularly for services

Support for Data Services

is not entirely straightforward. Data
services are becoming more complex than
traditional services, and scholarly
services take is to hire a data librarian,
to develop data curation services,
initial appointment, Cornell has
the job descriptions of additional
services, and the university also has
so that its focus is on science data
open-source tool infrastructure. When libraries
infrastructure locally, the work is
gainst institutional grant programs. Some of these projects are
for successful projects, transitioning
production-scale service and program
do not progress beyond the proof-
data services and developing infras-
be a part of a senior administrative
ful support to set direction and
ent. Because research data services
they often aren't fully integrated
At Cornell, primary respon-
the Chief Technology Strategist,
but because of the cross-functional nature of data curation, the Associate
University Librarians for Research and Learning Services and for Digital
Scholarship and Preservation Services also play important roles. The on-
going need for effective collaboration and communication across multiple
units and functional areas, regardless of the place of data services within
the organization, suggests that a cross-functional group such as CUL's Data
Executive Group is both useful and necessary.

Regardless of a library's organizational structure, staffing for data
services is a significant challenge. Ideally, staff in these roles should pos-
ssess some combination of research and disciplinary experience, as well as
information management and technology skills. Research experience and
disciplinary knowledge are particularly important for establishing cred-
ibility with researchers and for identifying and understanding their needs.
Library and information science programs are offering specializations, cer-
tificates, and continuing education programs to train new professionals and
to retool practicing librarians for the challenge, and this development holds
some promise for rounding out or upgrading staff skill sets. CUL has had
good success in attracting qualified applicants for positions with research
data responsibilities, and it has used the Data Discussion Group, an infor-
mal group with open membership, to attempt to broaden the conversation
about data and data services to include more staff.

Refining the Scope for Data Services

Libraries must also define the scope of services they wish to provide. With
more traditional publication outputs, it has long been sufficient for librar-
ies to limit their activities to acquiring and managing the finished prod-
acts of scholarship. Data management, on the other hand, is much more
effective when approached from a full research and life cycle perspective,
but libraries are not generally viewed as partners in the research process.
While this is partly a matter of whether or not library staff have the needed
skills, it also represents a fundamental change in the way libraries approach
the management of information. It implies a way of working with scholars
that is responsive and iterative, and that balances the norms of disciplinary
practices with library and information management best practices. Individ-
ual data librarians at CUL with sufficient disciplinary expertise have been
able to work with research groups at various stages in the research process,
but this work requires a great deal of time and specialized knowledge. While this work has resulted in new partnerships and the inclusion of the library in research grants, it is not something we've yet learned how to scale up to serve more than a small number of research groups.

A related issue is whether libraries should direct resources toward generic infrastructure that meets the needs of many researchers in a basic way, or toward more specialized or discipline-specific services, or both. Researchers are often more deeply aligned with their disciplines than their institutions and may prefer domain-based infrastructure to that of their home institution. They also frequently collaborate across institutional boundaries, making it complicated to determine where support is best provided. For the library, the prospect of providing specialized services to selected researchers collides with deeply rooted professional values of equal service and access to information for all. This is a particularly difficult problem if a service does not scale well to meet the needs of all who might wish to use it. The need for disciplinary expertise for some kinds of data services means that library administrators will be forced to decide which disciplines will be served more fully, presuming they are not in the position to staff for data services at the same level they already do for public services and collection development. CUL's approach so far has been to continue to develop and provide basic services and infrastructure that meet most of the basic needs of most researchers (primarily the institutional repository, eCommons, and consultations through the RDMSG, as well as developing the Datastar system), while offering more specialized services on a project-by-project basis, with grant funding or other financial support.

Many research libraries have embraced the work of curating the outputs of "small science" and are relatively comfortable with curating small-scale research data, but few have begun to address the challenges of larger-scale data. So-called "big data," datasets large enough to preclude the use of standard tools and protocols, is sometimes thought of as a solved problem because its curation is frequently inseparable from its collection, management, distribution, and use. Nevertheless, preservation of data once the project or facility that generated it is retired remains an issue. It is also likely that many datasets of more modest scale (but still too large to be handled well by an institutional repository) will require curation and preservation. Preservation is a function that falls fairly clearly within the scope of the library's work.
An Institutional Perspective on Data Curation Services

1e and specialized knowledge. While ship and the inclusion of the library we’ve yet learned how to scale up to search groups.

1 should direct resources toward ge- needs of many researchers in a basic discipline-specific services, or both. Re- al with their disciplines than their in- d infrastructure to that of their home- borate across institutional boundar- ne where support is best provided, iling specialized services to selected professional values of equal service is a particularly difficult problem if a needs of all who might wish to use it, some kinds of data services means ed to decide which disciplines will be not in the position to staff for data do for public services and collection has been to continue to develop and re that meet most of the basic needs institutional repository, eCommons, and well as developing the Datastar sys- services on a project-by-project basis, support.

braced the work of curating the out- ely comfortable with curating small- an to address the challenges of larg- datsets large enough to preclude the is sometimes thought of as a solved- ently inseparable from its collection, nevertheless, preservation of data once is retired remains an issue. It is also est scale (but still too large to be han- ) will require curation and preserv- falls fairly clearly within the scope of the library’s mission, yet the technical and staff capacity to deal with larger datasets may be found more readily in other units, such as central information technology or high-performance computing. Libraries are quite good at managing digital objects that are small enough to be easily downloaded via HTTP, but medium- to large-scale datasets require other means for access and distribution that most libraries do not currently support. Libraries may be able to ingest larger datasets into repository or archival systems for preservation, but they are not used to provisioning for access at larger scales. If we are to step up to the challenge of hosting larger datasets, we need to develop the capacity to support access as well, or find ways to collaborate with other campus service providers to accomplish this. At Cornell, a review of campus-wide IT support for research is underway, and we expect some useful recommendations to emerge from that exercise.

Dealing With Uncertainty

A final challenge is that of constantly reevaluating the evolving landscape, and there is an ongoing need to assess researchers’ requirements for data curation services. Having put significant effort into planning and establishing services, it is worth asking whether funder mandates and the library’s response have resulted in an increase in demand for data services, and where we should go from here. So far, beyond the demand for consultative services, we do not yet know how the demand for services will grow. The lags inherent in the review and award process, combined with the average duration of a research grant, mean that we have not yet seen the surge in demand for data curation services we expect to see as the first grant-funded projects subject to DMP requirements draw to a close. At Cornell, we know that the average length of time between proposal submission and notification of an award is 10 months. That means the very first awards with DMPs required in the grant proposals were likely made in the fall of 2011. With 85 percent of Cornell awards lasting three or more years, those first awards will not expire until fall of 2014 (Figure 3). This gives the RDMSG and allied service providers, including the library, a window of opportunity to learn from researchers’ expectations, as expressed in the DMPs we have had an opportunity to review, and to plan to fill service gaps that might exist.

More broadly, technology and hardware change rapidly, and modes of scholarly communication and research practice evolve. Librarians with
liaison responsibilities tend to make an effort to avoid over-contacting faculty to gather information, yet the necessity of doing so to understand faculty needs is ongoing. Library staff are challenged with staying current in developments in information technology, standards, and disciplinary practice. Solutions for data distribution and archival systems may emerge from multiple sectors besides academic libraries; these include major funding agencies, publishers, nonprofit organizations, professional societies, university and library consortia, and community efforts rooted in specific disciplines. It is not clear which of these efforts will emerge as successful and enduring, or whether we should continue to expect to have a whole host of players meeting different needs in different contexts. In spite of that uncertainty, the view at Cornell is that research data are an important part of the scholarly output of the institution, and that it is our responsibility to assume a stewardship role to the very best of our ability.

ACKNOWLEDGMENTS
I thank Anne Kenney, Wendy Kozlowski, and Sarah Wright for their thoughtful comments on an earlier version of this chapter. Without the members of all of the committees, working groups, and project teams described in this chapter, there would be nothing to write about.

REFERENCES
An Institutional Perspective on Data Curation Services


October 2011
First awards with DMPs

October 2014
First 3-year grants with DMPs