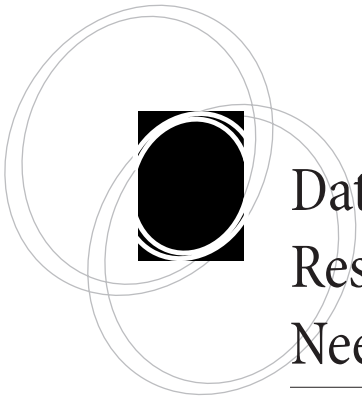

Data Curation: A Study of Researcher Practices and Needs

Merinda McLure, Allison V. Level, Catherine L. Cranston, Beth Oehlerts, and Mike Culbertson

abstract: Colorado State University librarians conducted five focus groups with thirty-one faculty, research scientists, and research associates. The groups explored: (1) The nature of data sets that these researchers create or maintain; (2) How participants manage their data; (3) Needs for support that the participants identify in relation to sharing, curating, and preserving their data; and (4) The feasibility of adapting the Purdue University Libraries' Data Curation Profiles Toolkit[1] interview protocol for use in focus groups with researchers. The authors report their review of related literature, themes that emerged from analysis of the focus groups, and implications for related library services.

Introduction and Research Questions

In spring 2012, our research team of five Colorado State University (CSU) librarians conducted five ninety-minute focus groups with a total of thirty-one self-selected faculty, research scientists, and research associates. The participants were collectively affiliated with seven of CSU's eight academic colleges or with the university administration. The purpose of this qualitative study was to explore the following research questions: (1) What is the nature of the data sets that these researchers currently create and maintain? (2) How are these researchers currently managing their data sets? (3) What needs for assistance and support do these researchers identify in relation to sharing, curating, and preserving their data sets? and (4) What is the feasibility of adapting the Purdue University Libraries' Data Curation Profiles (DCP) Toolkit[2] (hereafter, the Toolkit) protocol for use in focus groups with researchers? The research team created a focus group protocol by adapting the Toolkit protocol and used thematic, template analysis[3] to code and analyze the focus group discussions.

portal: Libraries and the Academy, Vol. 14, No. 2 (2014), pp. 139–164.

Background

The CSU Libraries provides critical support for the university's research, teaching, and outreach missions as a land-grant institution—that is, one which receives partial federal support dating to the Morrill Acts—and as a Carnegie Research University (Very High Research Activity), a university recognized by the Carnegie Foundation for the Advancement of Teaching for its advanced, active research programs. CSU employs 1,560 faculty and enrolls approximately 29,500 students.[4] In the university's fiscal year 2012, CSU researchers received a total of 2,049 research awards, valued at more than $267 million.[5] Major funding agencies support research at CSU, including the National Science Foundation (NSF), the National Institutes of Health (NIH), the U.S. Department of Defense (DoD), and the U.S. Department of Agriculture (USDA).

In 2010, CSU implemented the strategic vision determined by a 2009 task force when it merged the libraries and campus Academic Computing and Network Services (ACNS) into one organization. This reorganization is likely to benefit future efforts across campus to improve data curation—the organization, preservation, and management of data to ensure that they are retrievable for future research or reuse. ACNS and the CSU Libraries will be two primary internal stakeholders as campus data-curation efforts evolve. Since 2010, CSU Libraries administration and faculty have invested increased effort in exploring data management and the potential development of supporting services facilitated by the libraries and ACNS. The libraries has partnered in key campus programs related to data management. In 2011, it developed data-management plan templates to help CSU researchers author grant proposals that conform to the NSF Data Sharing Policy.[6] In 2011–2012, the libraries participated in the Association of Research Libraries/Digital Library Federation (ARL/DLF) E-Science Institute, a program to help libraries support research in e-science, large-scale scientific research carried out through global collaborations on the Internet.[7] CSU and the University of Colorado jointly support the Digital Collections of Colorado institutional repository, where researchers and other scholars can archive such materials as unpublished scholarly works and lectures.[8] The CSU Libraries is currently focused on collecting in the repository textual scholarship produced by researchers and students, and has begun to explore data hosting by taking in research data sets.

Literature Review

This literature review provides context for the developing engagement with the curation of research data by American institutions, including funding agencies, universities, and academic libraries. It addresses the significance of data repositories in relation to data curation and data sharing, the pertinence of academic libraries' participation in data curation, and new roles in data curation for librarians.

Path to the Present

Data curation is well defined by Sarah Shreeves and Melissa Cragin as "the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation

and organization of these data for access and use over time."[9] The now widespread preoccupation with data curation and data sharing, and the prevalence of related local, national, and international initiatives, are understood to be outcomes of the "growth of data-intensive research," the scale of which is underscored by Gordon Bell, Tony Hey, and Alex Szalay's 2009 characterization of "data-intensive science" as an emerging, fourth scientific research paradigm (following experimentation, theoretical science, and computer simulations).[10]

Anna Gold has comprehensively chronicled the development of data curation.[11] Liz Lyon has noted that influences include "an increasingly open scholarly communications agenda," "Web tools and applications which accelerate the 'publication' process," and "economic drivers for greater accountability and transparency, to show the impact of public investments in science."[12] The emergence and influence of open scholarly communication and open data agendas, internationally, are significant. Jennifer Molloy addresses the principle and development of open data, and Greg Tananbaum explains and provides the case for open data policy development and implementation by research funders.[13]

Recent articles by Neil Beagrie, Robert Beagrie, and Ian Rowlands; Jake Carlson; and Florian Diekmann identify many of the studies that have investigated the evolving data-curation needs of researchers in the United States and internationally.[14] Data management is critical to researchers throughout the life cycle of data.[15] The advantages of data sharing are many. Scholars may benefit from accessing and using data produced by others, from preserving and sharing their own data, and from the benefits to research and the public good which can derive from the sharing and reuse of data. Christine Borgman points out that data sharing makes it possible to reproduce, verify, advance, and publicly disseminate research.[16] Michael Whitlock suggests that scholars can conduct more thorough metanalyses, use data in teaching and learning, and reduce the risk of data loss by both publicly and locally archiving data.[17] Michael Witt notes that sharing supports the interdisciplinary use and repurposing of data.[18] Data sharing may even, Gail Steinhart argues, enable researchers to address "errors in data in response to feedback from users."[19] Borgman contends, "If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them."[20] Researchers, however, perceive and must negotiate a variety of barriers (technological, social, organizational, financial, and other) related to sharing their data.[21] Yi Shen and Virgil Varvel Jr. suggest that a primary measure of the success of data-management services may include researchers' "appreciation and implementation of data management in general."[22]

Data management is critical to researchers throughout the life cycle of data.

Data repositories provide one important means through which data may be curated and shared. Mark Parsons and his coauthors describe "a vision of discoverable, open, linked, useful, and safe collections of data," within a "data ecosystem" perspective.[23] Karen Baker and Lynn Yarmey point out that repositories "allow local data to be translated to the larger context of global environments

Data repositories provide one important means through which data may be curated and shared.

and multidisciplinary arenas."[24] Baker and Yarmey discuss "data stewardship" as the tending of multiple "related repositories" from a big-picture perspective. In this view, data are understood to move through a "web-of-repositories," acquiring a collectively defined, "cumulative value."[25] Shared data sets must be accurately identified and cited. Matthew Mayernik surveys current developments in this area.[26]

Data Curation and Libraries

The library and information science (LIS) profession has foreseen and responded to the emergence of new opportunities in data curation. The Association of College and Research Libraries (ACRL) Research Planning and Review Committee identified data curation as a top trend for academic libraries in 2012. The committee cited as recent drivers of the data-curation trend the increasingly common practice of scholarly journals publishing articles with accompanying data sets, as well as the NSF Data Sharing Policy requirement that data-management plans accompany grant proposals as appropriate.[27]

The Association of College and Research Libraries (ACRL) Research Planning and Review Committee identified data curation as a top trend for academic libraries in 2012.

Writing in 2010, Anna Gold predicted that the near-term horizon would see a handful of research libraries contributing to "national digital curation strategies," widespread academic library and librarian involvement in "the development of campus-based data curation strategies," and the growth of related, graduate-level and professional development programming for LIS students and librarians.[28] She suggested that while research libraries "are unlikely to be in a position to curate major collections of digital data," even with stable funding and local expertise, the hope and expectation is that they may contribute to "establishing collaborative networks of organizations that will be capable of executing this responsibility."[29] The technology infrastructure needed to support data curation means that campus information-technology units are key stakeholders, and common partners for libraries, in data-curation efforts. Surveys of services and reports and guides on evolving practices now offer examples and insights for libraries as they develop and evolve research data services.[30]

New Roles for Libraries and Librarians

The LIS literature increasingly documents academic libraries' evolving responses to data-curation needs and challenges. This literature affirms libraries and librarians as well-suited for work in this arena, given their long stewardship of collections and support of research and scholarly communication; their more recent implementation of institutional repositories; and their expertise, Gail Steinhart notes, in "archival practices, cataloging and indexing, development of platforms for discovery and distribution, and education and user support."[31]

Recent library and librarian activities in this arena demonstrate that librarians are discerning and responding to a wide spectrum of opportunities. Dianne Dietrich and her coauthors reviewed data-management and data-sharing policies of research funding

agencies. They identified "gaps between data management goals and implementation realities."[32] These gaps consequently suggest opportunities for libraries to assist researchers with identifying metadata and data standards and existing data repositories; with embargoing data when appropriate; and with building the infrastructure needed to demonstrate compliance with policy-based requirements for making data accessible.[33] Lyon outlines the array of research data management services that may be provided by libraries and considers that liaison librarians, with their primary responsibility for support of faculty and students, may be particularly well positioned to support research data management.[34] Tracy Gabridge agrees that liaison librarians can apply their expertise to new roles, such as the determination of the best repository for a given data set, consultation with researchers on appropriate standards and life-cycle planning for their data, and the instruction of students in prudent data-management practices.[35] Sarah Williams's work likewise suggests a role for librarians in helping researchers identify data sources and disseminate data through appropriate repositories.[36] Mark Newton, C. C. Miller, and Marianne Stowell Bracke identify new librarian roles in the four categories of "data identification, mediation, selection and appraisal," related to the local collection of researcher data to a data repository.[37] They suggest as critical the following librarian skills: the ability to encourage researchers to deposit their data in the institution's repository; fluency in translating the capabilities of the repository system for researchers; and perhaps most importantly, the ability to communicate and interact effectively with faculty.[38]

Research data-management and support needs challenge libraries and librarians, Liz Lyon says, to "re-position, re-profile and restructure to be fit for purpose in a data-centric research landscape."[39] Carol Tenopir and her coauthors report that in ARL libraries many librarians have professional interest in, and feel equipped for, future engagement in research data services.[40] LIS programs appear to be gradually responding to the need to prepare future librarians for new roles. Rebecca Harris-Pierce and Yan Quan Liu identified related, graduate-level course offerings by sixteen of the fifty-two North American LIS schools accredited by the American Library Association that they considered in 2012.[41] Nicholas Weber, Carole Palmer, and Tiffany Chao suggest that data curation will require individuals with "a set of combined competencies from domains like information science and computer science, as well as the natural sciences" and that "the future success of LIS curation programs will require new strategies for attracting promising students from across traditional campus departments."[42]

Researcher Needs and Library Responses

Our study adds to a growing body of literature that reports librarian investigations of researchers' data-curation needs and the development of related library services and initiatives. Leslie Delserone reports the convergence of hiring, research, program development, and (with other units) a campus scan of existing "computationally intensive research," that all supported the University of Minnesota Libraries' initial entry into data management.[43] A key finding of this scan was that researchers were keen to "relieve themselves of the day-to-day burden of administering data management solutions."[44]

Christie Peters and Anita Riley Dryden analyzed the data-management needs of principal investigators working on NSF and NIH grant-funded projects.[45] The authors

identified researcher needs for assistance including "help with the grant proposal process in general, especially assistance with funding agency data management requirements, help identifying campus data-related services, publication support, and targeted research assistance attendant to data management."[46]

Kathryn Lage, Barbara Losoff, and Jack Maness conducted interviews with University of Colorado Boulder faculty and graduate students in science disciplines. Through their analysis, the interviewers created eight researcher personae, each of which captures researchers' "range of attitudes and needs regarding the type of datasets created, existing data storage and maintenance support, disciplinary culture or personal feelings on data sharing, and receptivity to the library's role in data curation." The authors suggest "that librarians target researchers similar to five" of the eight personae, as these researchers are most likely to be receptive to working with the library on data-curation activities.[47]

Melissa Haendel, Nicole Vasilevsky, and Jacqueline Wirz describe challenges encountered by the NIH-funded eagle-i initiative, intended to support biomedical research by networking the information respositories of multiple academic institutions through the creation of publicly searchable records describing those repositories' research resources. The authors observe and discuss the limited use of formal inventory systems and metadata by many academic research laboratories, which could facilitate ready linking to related data and lab publications in and across data repositories. Haendel, Vasilevsky, and Wirz also describe the limited influence of national data-sharing developments on the use of data-management plans in laboratories and see a need for increased education and cultural change among scientists in relation to personal responsibility, good practice, and ethics in data management and data sharing. The authors suggest that librarians are positioned to extend their long interest in information literacy to data and data curation.[48]

As libraries continue to investigate and respond to researcher needs, the Toolkit provides librarians with one structured methodology for collecting the very specific, granular level of detail that is needed to understand the needs of individual researchers in "managing, sharing or curating their data" and to determine how to support these needs.[49] The Toolkit is a guide to help librarians and other information professionals identify the data needs of researchers by interviewing the researchers themselves. The Toolkit then provides instructions for creating a data-curation profile, "essentially an outline of the 'story' of a data set."[50] A completed profile describes the data set; tells how the researcher handles, manages, and shares it; and summarizes the researcher's needs for the data. Michael Witt and his coauthors detail the initial development of the semistructured interview and data-curation profile methodology that was subsequently formalized in the Toolkit.[51] Jake Carlson describes the workshops that have disseminated the Toolkit nationwide and been provided to librarians—such as us—to train them in its use.[52] Our study presents focus groups as an additional research methodology for libraries to use as they undertake to ascertain the practices and needs of local researchers and begin to address the development of supporting services.

The Toolkit is a guide to help librarians and other information professionals identify the data needs of researchers by interviewing the researchers themselves.

Procedures

We obtained CSU Institutional Review Board approval, and between February and April 2012 we recruited participants for, and conducted, five ninety-minute focus groups. Thirty-one self-selected CSU employees with faculty, research associate, or research scientist employment status attended. We created a focus group protocol (see Appendix) by adapting the Purdue University Toolkit's interview protocol.[53] We used focus groups to explore whether this methodology can provide librarians with the broad understanding of researchers' data-curation needs that is desirable in the early stages of developing related library services.

Recruitment

We recruited participants using template e-mails that included a link to the study's online registration form. The demographics of those who took part are reported in Table 1. More than 50 percent of the participants were faculty; more than 50 percent were female; and researchers affiliated with the College of Natural Resources comprised the highest percentage of participants (38.71 percent) affiliated with any one CSU college.

We scheduled five focus group sessions prior to recruiting participants. This decision was based on both our (correct) assumption that we would succeed in recruiting no more than fifty participants, as well as Richard Krueger and Mary Anne Casey's advice that "the ideal size of a focus group for most noncommercial topics is five to eight participants" and that focus groups of this kind should not exceed ten participants per group.[54]

Conduct of the Focus Groups

Prior to conducting the focus groups, we hired a social sciences faculty member, with specialized training in focus-group methodology, who provided a two-and-a-half hour training session to the research team. This training importantly addressed the research team members' lack of formal education in focus-group methodology and provided an opportunity to practice and receive expert feedback on use of the protocol (see Appendix).

Each focus group was attended by between four and nine participants and was moderated and co-moderated by two (rotating) research team members. The focus groups were digitally recorded and professionally transcribed verbatim. After each group departed, the moderator and co-moderator debriefed to note significant themes that surfaced in the discussion and any additional details that might later inform analysis. Krueger and Casey recommend this practice.[55]

Data Analysis

Prior to analysis, research team members compared the professional transcripts to the recordings to confirm the accuracy of the transcripts and to note any corrections. The principal investigator and co-investigator conducted a thematic analysis of the focus group transcripts using NVivo software; the thematic coding, template analysis technique defined by Nigel King; and the validation strategy of peer review and debriefing, to ensure the trustworthiness of the analysis.[56] In our analysis, we treated all five focus groups as one data set. We did not attempt to analyze differences between groups as

Table 1.
Participant Demographics

Participants	Number	Percentage
Employment Status		
Faculty	18	58.06
Research associate	8	25.81
Research scientist	5	16.13
Total	**31**	**100**
Gender		
Female	17	54.84
Male	14	45.16
Total	**31**	**100**
College Affiliation		
College of Natural Resources	12	38.71
College of Applied Human Sciences	4	12.90
College of Liberal Arts	4	12.90
College of Natural Sciences	4	12.90
College of Veterinary Medicine and Biomedical Sciences	3	9.68
College of Engineering	2	6.45
College of Agricultural Sciences	1	3.23
Administration	1	3.23
Total	**31**	**100**

we assigned members to each focus group according to availability and without regard to demographics.

In accordance with the template analysis technique, the principal investigator created an initial set of thematic codes based on her review of one transcript. The co-investigator then coded the same transcript with this initial set of codes and with new codes created by her as she worked. The investigators discussed this work, arriving at consensus concerning changes to the initial set of codes and together revising the template set of codes. We then independently coded each transcript, meeting between transcripts to review and adjust the codes—as needed and always based on consensus—before proceeding. In this inductive coding process, our template set of codes changed most between our coding of the first and second transcripts and remained relatively stable thereafter. Lastly, the investigators determined the smaller set of major themes evidenced by this work and grouped all codes under these themes. We report these themes and illustrative participant quotations in the following section.

Findings

Participant Research Projects

Participants were asked to characterize research in their disciplines and to describe one research project in which they had participated. They described qualitative, quantitative, and mixed-methods research. In alignment with the demographics of those who took part (see Table 1), most participants described research in the sciences or social sciences; fewer study members described research in the arts or humanities.

The majority of participants reported projects that involved original data collection. Only a handful reported projects that utilized previously collected or aggregated data. Project descriptions indicated the use of gene sequencing; human data; mass spectrometry; remote sensing; geographic information systems (GIS); radiotelemetry; radar; interviews; government archives and historic documents; surveys; satellite images; sound files; images or photographs; spatial data; integrated physical, ecological, and social data; maps; student test scores; and financial or economic models. Several comments illustrate the diverse spectrum of projects described and the pertinence of effective data management in research:

The majority of participants reported projects that involved original data collection.

> [Participant:][57] We have an NSF project in [country] where we're looking at how herders move across the landscape and are adapting to climate changes in terms of their migration patterns. So we're collecting physical, ecological, and social data to analyze and figure out the health of the landscape along with strategies that herders are adopting for movement patterns that are also linked to policy implications.

> [Participant:] I'm in the [department]. And some of the data we'll be collecting in the very near future is a lot of sequencing data. So we'll have to deal with how to share that data, how to collect that data, in a manageable format. At the same time, colleagues in the lab deal with a lot of time lapse, video imaging, and data files, using confocal microscopy. That tends to require a lot of large files and storage volume.

> [Participant:] So we have some unique data features in our lab. We have traditional scientific data collection with files, and maintaining those files for storage. And then there are a lot of community-driven governing principles, where you have to get that data into an executable format [so] that anybody with a certain freeware can access it and re-analyze it. So we have to be able to translate our data into some universal format. Then we have to upload it to a public database that doesn't have any restrictions on its access. And that's an interesting feature, to be able to publish a lot of our findings is to be able to make that data accessible to criticism.

Research Project Data Characteristics

Data File Formats

The majority of participants mentioned using a core group of file formats in their research. These formats included Word, Excel, and PowerPoint; comma-separated values (CSV) files, in which the values in a table are saved as lines of plain text with the value in each column separated from the next column's value by a comma; portable document format (PDF) files, which enable electronic documents to be distributed with the same layout, formatting, and images as in the original; and relational database applications including Access, Oracle, and others that use Structured Query Language (SQL), a computer programming language used to query, insert, and modify data. Many focus group members also mentioned using Joint Photographic Experts Group (JPEG) and Tagged Image File Format (TIFF), formats used to store digital images; Moving Picture Experts Group (MP3) or other sound files; and for statistics, the Statistical Analysis System (SAS) or Statistical Package for the Social Sciences (SPSS) files.

Some participants mentioned working with GIS shape files and other geospatial data, including satellite imagery. Only select members of the study reported working with the Hierarchical Data Format (HDF), the Network Common Data Form (NetCDF), and formats specific to instrumentation, X-ray crystallography, and proprietary software programs such as NVivo. While lab notebooks, field notebooks, paper index cards, and animal or plant specimens are not electronic file formats, these were also mentioned as mechanisms for collecting data. One participant noted, "They're going out and not only are they writing in their field notebooks, they're taking photos and they're bringing specimens of plants and animals. I mean that's data too. And it's a little harder to store. And there's no way for us to back that up."

Data File Sizes

We asked participants to speak to the size of the individual files that their research generates and to the sum size of all files that their research produces over the course of a single project. For the purpose of this discussion we presented the following definitions: *small* (data sets up to 200 gigabytes [GB]); *medium* (data sets 200 GB to 10 terabytes [TB]); and *large* (data sets more than 10 TB).

Most participants indicated that their individual project data files are typically small. The few researchers who described commonly working with large files are engaged in scientific research involving mass spectrometry, meteorological, satellite, spatial, or sequence data. Those who work with large files noted that the transfer of files within a lab's computer networks, across the campus network, or to geographically distant sites presents ongoing challenges. They also expressed concerns regarding file storage, data integrity, data backup, and data transfer. Virtually all participants indicated that they expect their file sizes and their file storage needs to increase in the future.

Virtually all participants indicated that they expect their file sizes and their file storage needs to increase in the future.

Data File Standards

A few participants from science disciplines mentioned specific standards that they utilize in their data projects, such as GRIdded Binary or General Regularly-Distributed Information in Binary form (GRIB), used to store forecast weather data, and Binary Universal Form for the Representation of meteorological data (BUFR). Several of those who took part in the study mentioned the importance of using or adhering to standards to share data successfully. A number of natural resources researchers mentioned contributing data to repositories and shared network projects involving the Global Biodiversity Information Facility.[58] One researcher discussed work with the Data Observation Network for Earth (DataONE), an NSF-funded initiative to share biological data among repositories. Participants also named other joint efforts such as the Taxonomic Database Working Group, Species 2000, and the Encyclopedia of Life.[59]

Participants mentioned their use of and need for standards and indicated that they often have questions about standards. They commented that when they have questions on standards, data organization, and metadata—that is, information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage data resources—they would benefit from access to expert assistance.

Data-Management Life Cycle

We provided participants with a data life-cycle model (Figure 1) and asked them to identify the stages that they felt to be most significant or a focus in their work. Two prominent themes emerged from this discussion.

Life Cycle Functions and Data Management

Plan: determine what data need to be created or collected; identify standards for data and metadata

Create: produce/acquire data for purposes; store and backup locally

Keep: organize and store data

Produce derived productions in support of research (e.g. data summaries, reports, publications)

Transfer: deposit data and derivatives, including data sets, metadata and scholarly communications and publications, where they will be kept, maintained, managed, and can be shared with others

Share data:
- with others at CSU
- with others outside of CSU
- by depositing data into an institutional repository or subject-data repository (such as ICPSR, Dryad, GenBank, or other)

Figure 1. Data life-cycle example provided to focus group participants. In constructing this life cycle, we incorporated information from the Association of Research Libraries/Digital Library Federation (ARL/DLF) E-Science Institute and the Toolkit.

First, most participants felt that the *plan, create, produce,* and *transfer* stages are of most importance to their projects. Only one researcher identified the *create* stage as most important: "Creating or collecting data, [you] can't really do anything unless you do this. After that all [stages] are equally necessary." Several of those who took part talked about how much time they invest—or should invest—in planning, while noting their experience that even the best-laid plans are vulnerable to change. Software, technology, and laboratory procedures all change over time, and participants commented that over the course of a project they may discover better methods.

The relationship between planning and sharing was the second theme that emerged from discussion of data-management life-cycle stages. Study members indicated their interest in sharing data and their awareness that it is important to plan for sharing during the earlier stages of the life cycle. One talked about successfully sharing data with the Global Invasive Species Information Network (GISIN)[60] and that the opportunity to share data globally compels this participant's research group to significantly invest in sharing and disseminating their data. Others noted that national projects can force researchers to record data in uniform ways, so as to facilitate successful data sharing.

National projects can force researchers to record data in uniform ways, so as to facilitate successful data sharing.

Participants provided insightful comments as to stages or processes present in their research projects that did not appear to them to be reflected in our life-cycle model. Several individuals perceived of the life cycle as less linear than the model, given that there may be several iterations of a given stage, or movement back and forth between stages, once a project is underway: "I don't look at this as linear. I can see where you go through steps, perhaps again and again . . . I've also started looking at this from more of a business model or the economics at producing the data and how much it costs from data collection to ready to share." One individual conceived of the life cycle as beginning prior to planning: "I would add some sort of collaborating or networking to even get to the planning table with people." Participants noted that planning may be preceded by a quality-control assessment of existing data that have been identified for use in a research project.

We asked participants if any stages of the life cycle present particular challenges or are stages with which they would like assistance. Several researchers mentioned that most stages of the life cycle are challenging. One noted, "We struggle with all of them to be honest," and another commented, "I'm working on a project with the group that is not prepared for data management at all . . . I wonder if there might be some training or groups like this that sit down and do best practices around data collection, data storing." Some participants indicated that they would like more help with the storage and transfer of data, including data migration, as well as assistance with securely transferring data from research sites back to CSU: "For instance, collecting social science data or wildlife data in [international location]. You carry a hard drive with you, a strong hard drive, and have it backed up on there and on your computer. But if your computer is swiped on the way home . . ."

Sharing of Data and Products

Audiences

Participants identified the audiences with whom they currently share either their data, or products resulting from their data (such as reports). The four major audiences that study members identified were educators, government entities (from municipal to federal to international), the public, and other specialists and researchers (from graduate students to national laboratories to international organizations). Some participants are involved with national or international initiatives to share data. At times, focus group members mentioned their desire to expand their data or product sharing to additional audiences, including "new" audiences such as humanities or social science researchers who are now utilizing GIS data.

Data and Products

Collectively, respondents indicated that the products of their research may be proprietary or made openly available to both specialist and public audiences, as in the case of published journal articles. They mentioned a wide range of products that they may produce for specific audiences, including algorithms, best practices manuals, books or chapters, conference proceedings, data analyses, data sets distilled for a specific audience, databases, grant proposals, journal articles, maps, new patented crop varieties, newsletters for teachers, presentations, public policy papers, reports, researcher interviews published in the media, software and accompanying documentation, statistical models, student theses and dissertations, and Web sites. One participant commented: "I think a lot of our data results in not just secondary data analysis, but tertiary and quite a few progeny from the initial, based on what a particular person's interest is. They might package it into an XML format or some sort of format that they can move into another analytical piece of software. And then from there, they might move it into another quantitative statistical software package, and then from there, into a report. So I think that a lot of our data has a lot of progeny." A number of participants noted the increasingly prevalent practice of submitting distilled data to a publisher to accompany a published journal article. Researchers clearly connected data-archiving and data-sharing requirements to the need for comprehensive data management.

Terms and Conditions

Participants indicated several key reasons why they may share data and products: mandate by a grant funder, a journal, or a government entity; necessity, to functionally accomplish collaborative work such as student theses or research projects involving peer researchers at other institutions; and for the benefit of a specific audience. They also indicated terms and conditions that they consider before sharing their data and may ask or expect of individuals who use these shared data. While some terms and conditions are mandated (by organizations, for example), others are important to individual researchers or constitute common practice within a research community. Some participants are able to share data or research products only after a certain period, due

to public or industry funding for example; or only with previously approved audiences; or not at all, due to the nature of the data or the products. Multiple researchers noted that before sharing their data they are careful to consider their compliance with institutional review board requirements for protecting the identity of their human subjects. Additionally, participants emphasized that it is important to them to report their research findings in a published journal article prior to sharing their data. Study members also noted that they expect recipients of their shared data to inform the original researcher before further sharing the data and to attribute the data source in any published work resulting from use of the data.

Participants indicated that, in some instances, terms and conditions related to shared data may be captured in formal or informal memorandums of agreement (MOAs). One person noted that individuals do not always comply with MOAs, and several commented that while it can be helpful to require data users to employ a standardized citation to attribute shared data, erroneous citations and noncompliance with an agreement to cite the source of shared data are persistent problems. Participants indicated that it is essential to track data users in relation to funded research: "To produce the data, the agency wants to see results. And those come in terms of, how many people use the data? How many publications were there? And then tracking down has been really, really a tough problem."

Data-Management Plans

Data-management plans (DMPs) help researchers plan, articulate, and execute data management, as well as comply with funder or agency requirements. We asked members of the focus groups if they had ever created a DMP. Some had never or only recently become aware of the concept of DMPs. Responses also revealed varied perspectives on what a DMP entails and whether it is only a formal plan or may also name procedural workflows that for many researchers are embedded in their research process: "It's embedded procedure, in a sense." Several participants indicated that they had contributed to or individually authored a DMP. Some had utilized the CSU Libraries' NSF DMP templates, one researcher had used the DMPTool[61] linked from the DataONE Web site, and a few participants indicated that they would appreciate assistance creating DMPs. Specifically, some study members indicated that they were not aware of the metadata standards in their discipline and how they should incorporate relevant metadata standards in DMPs and in their work. One participant commented that they had primarily worked with DMPs that were related only to the transfer of data. Another viewed as a DMP their center's internal file management and record retention plan. The researchers who had authored DMPs to meet agency requirements most often indicated that they had submitted a DMP to adhere to the NSF Data Sharing Policy. Participants also named agencies, including the DoD, the USDA, and the

Data-management plans (DMPs) help researchers plan, articulate, and execute data management, as well as comply with funder or agency requirements.

NIH, as requiring some data-management details. Two participants indicated experience reading data-management plans while serving as proposal reviewers. One researcher noted contributing to DMPs included in proposals and interest in later learning from reviewers' comments about these DMPs. This individual perceives that the NSF is looking for the research community to define and refine best practices in data management.

Library Support

Education, DMP Templates, and Expert Consultation

Participants expressed interest in future training opportunities, for themselves and for graduate students, focused on the digital collection of data (as opposed to the continued use of paper lab notebooks, for example); managing data; new methodologies for recording data; and data-organization approaches and tools. One noted that while some researchers may have become more aware of conscientiously planning for data management, due to federal mandates, many younger faculty are seeing the value of planned data management early in their careers and all researchers would benefit from experts "helping us do it right the first time." Several participants agreed that it would be helpful to facilitate the sharing and documentation of campus researchers' experience and expertise.

One noted that … all researchers would benefit from experts "helping us do it right the first time."

Some participants expressed awareness or prior use of, and appreciation for, the CSU Libraries' data-management plan templates.[62] One focus group member suggested that it would be helpful if less specific or additional data-management plan templates were available. A number expressed interest in the possibility of receiving individual, expert feedback on the accuracy of their draft plans, prior to proposal submission and particularly in relation to plan content addressing data storage and security: "It would be nice if it wasn't just the NSF template. And it would be dreamy if we could—have somebody make sure that what we're saying is accurate once we interpret our needs, so that we're submitting accurate information into our agencies."

Storage

Participants had strong opinions on the topic of data storage, expressing interest in the potential benefits of more centralized campus data storage as well as concerns about the possibility of a concomitant loss of current individual or research unit control. One study member saw a clear distinction between providing support and services versus instituting standards and requirements, preferring the former as—in the words of another participant—"more library like." Another participant noted that they would value receiving guidance in designing policy to standardize their lab's data storage procedures for the lab's servers.

Some participants felt that ACNS should be able to purchase storage space and manage data storage more cost-effectively than individual research or academic units,

and perceived that many individual campus units and researchers are "battling" with managing data security. One participant commented: "It's very easy to see how having a central, university wide, storage and dissemination system for data would be much more cost effective, and probably better executed, than anything we could do ourselves. We're not computer scientists. We're not bioinformaticians. We're biologists and chemists. We're hacks when it comes to the computer work."

"It's very easy to see how having a central, university wide, storage and dissemination system for data would be much more cost effective, and probably better executed, than anything we could do ourselves."

At the same time, participants expressed concern about the ability of ACNS to facilitate both security and flexibility within a more centralized storage model that might not be suitable for all researchers' data. Some focus group members spoke to the potential power of centralization to assist the sharing of data across disciplines and researchers. In one conversation, participants speculated on the ability of computing technology to "match up different data sets with relevant variables," for example, and thereby assist new research based on existing data.

Dissemination and Discoverability

Participants appeared to be optimistic that the CSU Libraries could play a valuable role in enhancing the discoverability and dissemination of their data and research products. Some indicated that they or their unit are already using the institutional repository to host documents.[63] One participant indicated that they perceived the repository as facilitating not only data and document management but also accessibility and preservation. Another researcher envisioned the potential of subject-specific repositories or data dissemination facilitated by the libraries: "I think in GIS in particular, there's [*sic*] certain agency wide repositories. But, as was discussed, sometimes there might be benefits to having a CSU-wide tool that disseminates geospatial data."

Discussion also associated the CSU Libraries and its personnel with metadata expertise and indicated a number of participants' awareness of, and involvement with, metadata concerns. One participant envisioned that the libraries' future involvement in hosting data, and the libraries' possible addition of metadata as part of this process, might go some way in qualifying the reliability of data made available by the libraries. Another researcher noted, "We're overworked and trying to keep up. And so my head sinks because I'm just like, gosh, I want to be able to do a better job at documenting these data when we do go to share them. And providing that contextual information in formal metadata. We're just thin on time and resources to get that work done. So we would welcome help in that regard."

Participants commented that the libraries could help make all campus researchers more aware of one another's research data and, as a result, help them identify CSU colleagues to collaborate on grant proposals. Often comments in this vein noted that researchers' heightened awareness of one another's work, or more effective means for easily identifying one another's work, could help everyone avoid duplicative effort: "I do

believe that there's got to be some role for the library in solving this efficiency issue and making sure that data from different disciplines is more shareable in an interdisciplinary fashion." Other participants also emphasized the importance of interdisciplinary access to data: "And our role, particularly in environmental history, is to draw from data sets from other disciplines and to synthesize that information into narrative format for a more general audience. And that's very difficult for us to do if we don't have some kind of translation service there that allows us to see what's there and, qualitatively, what it's attempting to discover." One researcher spoke to the importance of determining how to facilitate the interdisciplinary use of data while avoiding a need for effort-intensive, case-by-case data customization: "If we intend on going into cross-interdisciplinary activities, then we're going to need common GIS databases where these other disciplines can tap in, and then that means, on our side, we need to learn what that standard is that people can commonly share—That community has to come up and say what their standards are going to be."

Digitization

One participant pondered whether the libraries could assist with data development and projects that—with the digitization of materials, for example—might bring together geospatial and historical data: "For instance, there's a map of historic trails in the State and it's on paper. We would be able to use it for all kinds of analyses if we had a digital version of that. But we've never had a project where we could justify doing it."

Another participant commented that researchers have limited space for storing the handwritten notebooks containing the work of their students. The individual noted that the value of work contained in these notebooks may become evident only after several years and so digitization of the notebooks could assist both space issues and preservation.

Discussion

The focus group discussions and our analysis provided valuable insights into each of our research questions. Our first question was (1) What is the nature of the data sets that these researchers currently create and maintain? Our protocol elicited detailed descriptions of diverse research projects; indication of the prevalence of original data creation or collection, as opposed to the reuse of data created or collected by others; indication of specific data file formats that researchers appear to commonly use; and indication of both the prevalence and predominance of large numbers of small (200 GB or less) data files in many research projects. These findings will inform the libraries' continued consideration of both data-set hosting in the Digital Collections of Colorado institutional repository and broader support for data services.

Our second question was (2) How are these researchers currently managing their data sets? Our protocol was somewhat less effective in supporting our exploration of this question because participants demonstrated widely varying familiarity with the concept and language of data life cycles and data curation, and varying awareness of formal data-management plans. Before using our focus group protocol again, we would consider how we might revise the protocol in relation to this issue. It was very valuable, however, to confirm these variations in familiarity and awareness and that, as a result,

outreach and education efforts in these areas would be a relevant focus. We also learned how researchers may conceptualize the data life cycle and its stages; that this sample of researchers is, overall, open to the concept of data sharing and is aware that data sharing is becoming an imperative that they should expect and plan for; that the proprietary nature of some data, and other factors, preclude data sharing by some researchers; that some participants are actively sharing data and are aware of issues relating to the compatibility of data standards; and that data-sharing audiences and venues can be research- and researcher-specific.

Participants demonstrated widely varying familiarity with the concept and language of data life cycles and data curation, and varying awareness of formal data-management plans.

Our third question was (3) What needs for assistance and support do these researchers identify, in relation to sharing, curating, and preserving their data sets? Our focus group protocol, and our participants, exceeded our expectations with regard to this question. Members of the focus groups readily suggested multiple aspects of data curation and data sharing with which they would like assistance and often openly described related, self-perceived deficits in skill and knowledge. Notably, they did not generally define whether the libraries, ACNS, or another campus body could best address their needs, nor did they suggest that the libraries would not be a suitable entity to address their needs. Rather, they appear to be most concerned with receiving the expert assistance that they need and desire, and generally unconcerned with which campus entities might provide this support. This lack of concern is significant, suggesting that many researchers are likely to be amenable to library services supporting data curation, provided that these services appropriately address researcher needs.

Members of the focus groups readily suggested multiple aspects of data curation and data sharing with which they would like assistance.

Our fourth question was (4) What is the feasibility of adapting the Purdue University Libraries' Data Curation Profiles (DCP) Toolkit interview protocol for use in focus groups with researchers? We believe that our protocol and focus groups were effective and appropriate to the libraries' current needs at this early stage in the libraries' consideration of future supporting services. While individual interviews would certainly yield much more individually specific, granular detail to assist data-curation support of specific, individual researchers, focus groups proved to be a more time- and energy-efficient method for us to gain rich insight into both local commonalities and variations in researcher behaviors, needs, and perspectives.

Limitations

Several aspects of our study limit the transferability of our results to our wider campus community and to the researcher populations of other higher education institutions. A few of these limitations might be mitigated by the revised design of future studies.

First, we did not use a specific, recognized sampling technique, such as maximal variation sampling, in which respondents are chosen to be as different as possible from one another.[64] We recruited widely and comprehensively by advertising the study to all campus faculty, research scientists, and research associates and were pleased to succeed in recruiting a discipline-diverse sample of thirty-one participants. We accurately anticipated when designing our study that we would have difficulty recruiting a discipline-diverse sample large enough for us to cluster participants in discipline-specific focus groups, and it was important to us to gain insights into the perspectives of researchers affiliated with a range of disciplines. With additional personnel resources to invest in the conduct and analysis of focus groups, future studies might succeed in enrolling enough participant researchers to construct appropriately sized, disciplinary-specific focus groups. These discipline-specific groups might yield richer insights into disciplinary differences that can inform library data curation.

Second, we used focus groups as our sole methodology, rather than a triangulation of data-collection methods.[65] Our experience suggests that in a similar, future study, it would be feasible for focus group participants to also complete a valid, reliable questionnaire to collect straightforward details, such as file types that researchers use. The use of a questionnaire would increase focus group discussion time for questions that specifically benefit from the dynamic dialogue that can be facilitated in groups. It is also reasonable to expect that the combination of data that could be collected through both a questionnaire and focus groups could strengthen the validity of the findings. We agree with the perspectives of other researchers that individual interviews may be preferable to methodologies such as focus groups for understanding researcher data-curation needs at the very granular level.[66] In designing this study, however, we made a decision to use focus groups—rather than individual interviews—given the CSU Libraries' current stage of data-curation considerations, the limited personnel resources that could be invested in this research, and our interest in testing the feasibility of this approach.

Third, while we employed careful, internal peer review by the principal investigator and co-investigator as we coded and analyzed our data, we did not use additional validation strategies that are established for use in qualitative research.[67] It is reasonable to expect that the integration of additional validation strategies could increase the validity of a similar, future study's findings.

Implications

Our findings have several key implications for the CSU Libraries' future engagement with data curation that may also be relevant for libraries and librarians elsewhere that are beginning to explore their roles in data curation. First, it appears that many researchers are amenable to receiving expert assistance with multiple aspects of data curation and sharing, particularly given their own limited resources, their related and self-perceived skill and knowledge deficiencies, and the many other demands on their time and energy. These researchers are foremost concerned with receiving needed, quality assistance and do not necessarily have preconceived perceptions that the libraries are poorly equipped to offer assistance. As regards metadata, for example, some researchers perceive that librarians may be more expertly equipped than researchers to address this element of

data curation and sharing. The libraries might do well (as Lage, Losoff, and Maness suggest) to attempt to identify and work first with individuals who are open to assistance from the libraries.[68]

Second, the libraries and ACNS will need to determine practical and incremental priorities for supporting researchers. The libraries' current development of supporting services is still emerging, and our findings provide a starting point for prioritization by suggesting themes such as the prevalence of small (200 GB or less) data files versus more select instances of "big data"; the widespread use of a core set of common file formats; researchers' interest in facilitated connection, communication, and best-practices sharing with other researchers across campus; researchers' need to prepare select data sets to accompany published journal articles; and researcher interests in developing mechanisms to address issues such as the more consistent citation and attribution of their data.

These researchers are foremost concerned with receiving needed, quality assistance and do not necessarily have preconceived perceptions that the libraries are poorly equipped to offer assistance.

Third, both our literature review and our study demonstrate that while librarians undertake to develop new skills and knowledge appropriate to new roles in data curation, they may also leverage their existing skills and expertise to pursue education and outreach efforts. Such efforts may immediately and positively begin to support researchers' connection and communication with other campus researchers; promote more widespread awareness and use of the libraries' data-management plan templates and repository; enhance understanding of the data life cycle and data-management considerations associated with each stage of the life cycle; aid identification of existing local, regional, national, or international repositories for data sharing; facilitate identification and use of appropriate metadata standards; and encourage preplanning for sharing data files to accompany journal manuscripts. These goals are likely to be relevant, practical starting points for other libraries, also.

Finally, even as research continues to explore the data-curation needs of researchers, it is apparent that local studies can inform libraries and librarians about the behaviors, needs, interests, and concerns of researchers at individual institutions. It will be useful for libraries to assess the impact of the supporting services that they implement, over time. Our study suggests that libraries may find it practical and beneficial to use focus groups as they consider how to support researchers in managing their data, when broad but rich insights into researcher needs are helpful, and limited resources inhibit the use of more time- and cost-intensive methods, such as interviews.

Acknowledgements

We would like to thank the CSU Libraries Administration for funding the study costs; Jake Carlson, associate professor of library science and data services specialist at Purdue University Libraries; CSU faculty colleagues Jennifer L. Matheson, associate professor

of human development and family studies, and James Banning, professor in the School of Education, for their generous advice at multiple stages of this project; Donald Zimmerman, CSU emeritus faculty; Shu Liu, formerly of Colorado State University Libraries and now metadata and digital resources librarian at the University of California, Irvine; and our study participants and reviewers.

Merinda McLure is an associate professor and health and human sciences librarian at Colorado State University (CSU) Libraries in Fort Collins; she may be reached by e-mail at: merinda.mclure@colostate.edu.

Allison V. Level is a professor and coordinator for collections at CSU Libraries; she may be reached at: allison.level@colostate.edu.

Catherine L. Cranston is an associate professor and instruction librarian at CSU Libraries; she may be reached at: cathy.cranston@colostate.edu.

Beth Oehlerts is an associate professor and digital services librarian at CSU Libraries; she may be reached at: beth.oehlerts@colostate.edu.

Mike Culbertson is an associate professor, engineering librarian, and government documents librarian at CSU Libraries; he may be reached at: michael.culbertson@colostate.edu.

Appendix

Focus Group Questions

1. Please briefly describe the nature of data—qualitative or quantitative—that is collected or produced in your discipline or core research area.
2. Please think of one of your research projects that you consider to be representative for you. Tell us about your work with data in the context of this project. For example, did you analyze data produced by others or did you collect data in the field?
3. Still thinking of the research project that is representative for you, please tell us about the number of files, file sizes, and file formats of this project's data. Some examples of file sizes and formats are provided on page 1 of the handout.
4. Still thinking of the research project that is representative for you, please tell us about products produced from this data, such as journal articles or reports.
5. You have a handout that shows a data-management life cycle example. Please describe the parts of the data-management life cycle that are a focus for you.
6. Please describe how the data-management life cycle for your project differs from this one, if at all?
7. Are there parts of the life cycle where you would like help or more assistance?
8. Do you currently share your data and if so, who is your primary audience?
9. Are there any conditions that you consider prerequisite to sharing your data?

10. Have you ever created a data-management plan?
11. For those of you who have created a data-management plan, was this plan part of an agency mandate?
12. Have you been part of an agency review panel or process that looked at data-management plans that were submitted with the grant applications and if so, please tell us about your experience?
13. How do you see university libraries supporting your data management, if at all?
14. Before we end, do you have any last comments that you would like to share?

Notes

1. The workshop and the Data Curation Profiles Toolkit that is its focus are outcomes of collaborative research undertaken by the Purdue University Libraries and the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, which aimed to investigate researchers' data-management and curation practices and researchers' willingness to share their data. Purdue University Libraries, "Data Curation Profiles Toolkit," accessed February 13, 2013, http://datacurationprofiles.org.
2. Ibid.
3. Template analysis is a thematic analysis style and technique where the researcher develops a "coding template, usually on the basis of a subset of the data, which is then applied to further data, revised and reapplied." Nigel King, "Doing Template Analysis," in *Qualitative Organizational Research: Core Methods and Current Challenges,* ed. Gillian Symon and Catherine Cassell (Thousand Oaks, CA: SAGE, 2012), 426–27.
4. Institutional Research, Colorado State University (CSU), *Institutional Profile* (Fort Collins, CO: Institutional Research, CSU, 2012), 1, http://www.ir.colostate.edu/pdf/profile/profile_12.pdf.
5. Institutional Research, CSU, *2012–2013 Fact Book* (Fort Collins, CO: Institutional Research, CSU, 2012), 249–50, http://www.ir.colostate.edu/pdf/fbk/1213/2012_13_Fact_Book.pdf.
6. CSU Libraries, "NSF Data Management Plans," last modified January 12, 2011, http://lib.colostate.edu/repository/nsf; National Science Foundation, "Dissemination and Sharing of Research Results," accessed August 6, 2012, http://www.nsf.gov/bfa/dias/policy/dmp.jsp.
7. Association of Research Libraries/Digital Library Federation (ARL/DLF), "E-Science Institute Sponsored by ARL/DLF," *Association of Research Libraries/Digital Library Federation (ARL/DLF),* last modified July 11, 2012, http://uchc.libguides.com/content.php?pid=228797.
8. Colorado State University, the University of Colorado, Colorado School of Mines, and Colorado Mesa University, "Digital Collections of Colorado" (2013), http://digitool.library.colostate.edu/R/.
9. Sarah L. Shreeves and Melissa H. Cragin, "Introduction: Institutional Repositories: Current State and Future," *Library Trends* 57, 2 (2008): 93, doi:10.1353/lib.0.0037.
10. John Wood, "Coping with the Data Deluge," in *The Future of Scholarly Communication,* ed. Deborah Shorley and Michael Jubb (London: Facet, 2013); Liz Lyon, "The Informatics Transform: Re-Engineering Libraries for the Data Decade," *International Journal of Digital Curation* 7, 1 (2012): 127, doi:10.2218/ijdc.v7i1.220; Gordon Bell, Tony Hey, and Alex Szalay, "Beyond the Data Deluge," *Science* 323, 5919 (2009): 1297, doi:10.1126/science.1170411.
11. Anna Gold, *Data Curation and Libraries: Short-Term Developments, Long-Term Prospects* (2010), 1–33, http://digitalcommons.calpoly.edu/lib_dean/27.
12. Lyon, "The Informatics Transform," 128.

13. Jennifer C. Molloy, "The Open Knowledge Foundation: Open Data Means Better Science," *PLoS Biology* 9, 12 (2011), e1001195, doi:10.1371/journal.pbio.1001195; Greg Tananbaum, *Implementing an Open Data Policy* (Washington, DC: Scholarly Publishing and Academic Resources Coalition, 2013), http://www.sparc.arl.org/sites/default/files/sparc-open-data-primer-final.pdf.
14. Neil Beagrie, Robert Beagrie, and Ian Rowlands, "Research Data Preservation and Access: The Views of Researchers," *Ariadne: A Web & Print Magazine of Internet Issues for Librarians & Information Specialists* 30, 60 (2009), http://www.ariadne.ac.uk/issue60/beagrie-et-al; Jake Carlson, "Demystifying the Data Interview: Developing a Foundation for Reference Librarians to Talk with Researchers About Their Data," *Reference Services Review* 40, 1 (2012): 7–23, doi:10.1108/00907321211203603; Florian Diekmann, "Data Practices of Agricultural Scientists: Results from an Exploratory Study," *Journal of Agricultural & Food Information* 13, 1 (2012): 14–34, doi:10.1080/10496505.2012.636005.
15. Melissa A. Haendel, Nicole A. Vasilevsky, and Jacqueline A. Wirz, "Dealing with Data: A Case Study on Information and Data Management Literacy," PLOS [Public Library of Science] *Biology* 10, 5 (2012): 1, doi:10.1371/journal.pbio.1001339.
16. Christine L. Borgman, "The Conundrum of Sharing Research Data," *Journal of the American Society for Information Science and Technology* 63, 6 (2012): 1066–72, doi:10.1002/asi.22634.
17. Michael C. Whitlock, "Data Archiving in Ecology and Evolution: Best Practices," *Trends in Ecology and Evolution* 26, 2 (2011): 62, doi:10.1016/j.tree.2010.11.006.
18. Michael Witt, "Institutional Repositories and Research Data Curation in a Distributed Environment," *Library Trends* 57, 2 (2008): 193, doi:10.1353/lib.0.0029.
19. Gail Steinhart, "Libraries as Distributors of Geospatial Data: Data Management Policies as Tools for Managing Partnerships," *Library Trends* 55, 2 (2006): 265, doi:10.1353/lib.2006.0063.
20. Borgman, "The Conundrum of Sharing Research Data," 1059.
21. Djoko Sigit Sayogo and Theresa A. Pardo, "Exploring the Determinants of Scientific Data Sharing: Understanding the Motivation to Publish Research Data," in *Government Information Quarterly* 30, Supplement 1 (2013), ed. Marijn Janssen and Elsa Estevez, S19–S31, doi:10.1016/j.giq.2012.06.011.
22. Yi Shen and Virgil E. Varvel Jr., "Developing Data Management Services at the Johns Hopkins University," *Journal of Academic Librarianship* 30 (2013): 4, doi:10.1016/j.acalib.2013.06.002.
23. Mark A. Parsons, Øystein Goday, Ellsworth LeDrew, Taco F. de Bruin, Bruno Danis, Scott Tomlinson, and David Carlson, "A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science," *Journal of Information Science* 37, 6 (2011): 555–57, doi:10.1177/0165551511412705.
24. Karen S. Baker and Lynn Yarmey, "Data Stewardship: Environmental Data Curation and a Web-of-Repositories," Remote Repositories—Distant Origin, *International Journal of Digital Curation* 4, 2 (2009): para. 2.
25. Baker and Yarmey, "Data Stewardship."
26. Matthew S. Mayernik, "Data Citation Initiatives and Issues," *Bulletin of the American Society for Information Science & Technology* 38, 5 (2012): 23–28.
27. Association of College and Research Libraries (ACRL) Research Planning and Review Committee, "2012 Top Ten Trends in Academic Libraries: A Review of the Trends and Issues Affecting Academic Libraries in Higher Education," *College and Research Libraries News* 73, 6 (2012): 312, http://crln.acrl.org.
28. Gold, *Data Curation and Libraries.*
29. Ibid., 11–12.
30. Sayogo and Pardo, "Exploring the Determinants of Scientific Data Sharing"; Sheila Corrall, "Roles and Responsibilities: Libraries, Librarians and Data," chap. 6 in *Managing Research Data*, ed. Graham Pryor (London: Facet, 2012); David Fearon Jr., Betsy Gunia, Barbara E. Pralle, Sherry Lake, and Andrew L. Sallans, *Spec Kit 334: Research Data Management Services*

(Washington, DC: Association of Research Libraries, 2013); Shen and Varvel, "Developing Data Management Services at the Johns Hopkins University"; Sarah Jones, Graham Pryor, and Angus Whyte, *How to Develop Research Data Management Services—A Guide for HEIs* (Edinburgh: Digital Curation Centre, 2013), http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services.

31. Marianne Stowell Bracke, "Emerging Data Curation Roles for Librarians: A Case Study of Agricultural Data," *Journal of Agricultural & Food Information* 12, 1 (2011): 66, doi:10.1080/10496505.2011.539158; Lyon, "The Informatics Transform," 127; Tyler O. Walters, "Data Curation Program Development in U.S. Universities: The Georgia Institute of Technology Example," *International Journal of Digital Curation* 4, 3 (2009): 84, doi:10.2218/ijdc.v4i3.116; Steinhart, "Libraries as Distributors of Geospatial Data," 267.
32. Dianne Dietrich, Trisha Adamus, Alison Miner, and Gail Steinhart, "De-Mystifying the Data Management Requirements of Research Funders," *Issues in Science & Technology Librarianship* 70 (2012): discussion section, doi:10.5062/f44m92g2.
33. Ibid.
34. Lyon, "The Informatics Transform," 132–34.
35. Tracy Gabridge, "The Last Mile: Liaison Roles in Curating Science and Engineering Research Data," *Research Library Issues: A Bimonthly Report from ARL* [Association of Research Libraries], *CNI* [Coalition of Networked Information], *and SPARC* [Scholarly Publishing and Academic Resources Coalition] 265 (2009): 17, http://publications.arl.org/rli265/.
36. Sarah C. Williams, "Data Practices in the Crop Sciences: A Review of Selected Faculty Publications," *Journal of Agricultural & Food Information* 13, 4 (2012): 308–325, doi:10.1080/10496505.2012.717846.
37. Mark P. Newton, C. C. Miller, and Marianne Stowell Bracke, "Librarian Roles in Institutional Repository Data Set Collecting: Outcomes of a Research Library Task Force," *Collection Management* 36, 1 (2011): 56, doi:10.1080/01462679.2011.530546.
38. Ibid., 62–64.
39. Lyon, "The Informatics Transform," 128.
40. Carol Tenopir, Robert J. Sandusky, Suzie Allard, and Ben Birch, "Academic Librarians and Research Data Services: Preparation and Attitudes," *IFLA* [International Federation of Library Associations and Institutions] *Journal* 39, 1 (2013): 76–77, doi:10.1177/0340035212473089.
41. Rebecca L. Harris-Pierce and Yan Quan Liu, "Is Data Curation Education at Library and Information Science Schools in North America Adequate?" *New Library World* 113, 11/12 (2012): 598–613, doi:10.1108/03074801211282957.
42. Nicholas M. Weber, Carole L. Palmer, and Tiffany C. Chao, "Current Trends and Future Directions in Data Curation Research and Education," *Journal of Web Librarianship* 6, 4 (2012): 310, doi:10.1080/19322909.2012.730358; Ibid., 311.
43. Leslie M. Delserone, "At the Watershed: Preparing for Research Data Management and Stewardship at the University of Minnesota Libraries," *Library Trends* 57, 2 (2008): 206, doi:10.1353/lib.0.0032; Ibid., 202–10.
44. University of Minnesota Research Cyberinfrastructure Alliance, *Assessing Research Cyber Infrastructure Needs at the University of Minnesota* (2008), www.cni.org/tfms/2008a.spring/abstracts/handouts/CNI_Assessing_Butler.pdf., quoted in Delserone, "At the Watershed," 207.
45. Christie Peters and Anita Riley Dryden, "Assessing the Academic Library's Role in Campus-Wide Research Data Management: A First Step at the University of Houston," *Science & Technology Libraries* 30, 4 (2011): 389, doi:10.1080/0194262x.2011.626340.
46. Ibid., 397.
47. Kathryn Lage, Barbara Losoff, and Jack Maness, "Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder," *portal: Libraries and the Academy* 11, 4 (2011): 916, doi:10.1353/pla.2011.0049; ibid., 932.

48. Haendel, Vasilevsky, and Wirz, "Dealing with Data."
49. Purdue University Libraries, "Data Curation Profiles Toolkit"; Carlson, "Demystifying the Data Interview," 8.
50. Purdue University Libraries, "Data Curation Profiles Toolkit."
51. Michael Witt, Jacob Carlson, D. Scott Brandt, and Melissa H. Cragin, "Constructing Data Curation Profiles," *International Journal of Digital Curation* 4, 3 (2009): 93–103, doi:10.2218/ijdc.v4i3.117.
52. Carlson, "Demystifying the Data Interview."
53. Purdue University Libraries, "Data Curation Profiles Toolkit."
54. Richard A. Krueger and Mary Anne Casey, *Focus Groups: A Practical Guide for Applied Research* (Los Angeles: SAGE, 2009): 67.
55. Ibid., 94.
56. See note 3 for a definition of template analysis. King, "Doing Template Analysis," 426–27. Peer review and debriefing is a qualitative research validation strategy that is described in John W. Creswell, "Standards of Validation and Evaluation," chap. 10 in *Qualitative Inquiry & Research Design: Choosing Among Five Approaches* (Los Angeles: SAGE, 2013), 251.
57. Where we provide multiple quotations in series, we begin a quotation with "[Participant:]" to indicate that the quotation text following belongs to a different participant than the quotation text preceding.
58. Global Biodiversity Information Facility, "Global Biodiversity Information Facility," accessed August 6, 2012, http://www.gbif.org/.
59. "Data Observation Network for Earth (DataONE)," University of New Mexico, accessed August 3, 2012, http://www.dataone.org/; "Biodiversity Information Standards (TDWG), also known as the Taxonomic Databases Working Group, is a not-for-profit scientific and educational association that is affiliated with the International Union of Biological Sciences. TDWG was formed to establish international collaboration among biological database projects. TDWG promoted the wider and more effective dissemination of information about the world's heritage of biological organisms for the benefit of the world at large. Biodiversity Information Standards (TDWG) now focuses on the development of standards for the exchange of biological/biodiversity data." "About Us," Biodiversity Information Standards, TDWG (2011), http://www.tdwg.org/; "Species 2000 is a 'federation' of database organisations working closely with users, taxonomists and sponsoring agencies . . . The goal of the Species 2000 project is to create a validated checklist of all the world's species (plants, animals, fungi and microbes). This is being achieved by bringing together an array of global species databases covering each of the major groups of organisms." "About Species 2000," Species 2000 (2012), http://www.sp2000.org/; "Encyclopedia of Life," accessed August 27, 2013, http://eol.org/.
60. The Global Invasive Species Information Network (GISIN) is a repository for "aggregating and disseminating invasive species data in a standardized way." "Global Invasive Species Information Network (GISIN)," Natural Resource Ecology Laboratory, CSU, last modified February 1, 2013, http://www.gisin.org.
61. Regents of the University of California, "DMPTool," accessed August 6, 2012, https://dmp.cdlib.org/.
62. CSU Libraries, "NSF Data Management Plans."
63. In the time since we conducted our focus groups, the CSU Libraries' digital repository has been rebranded as the multi-institutional Digital Collections of Colorado, http://digitool.library.colostate.edu/R.
64. Maximal variation sampling is defined as "a purposeful sampling strategy in which the researcher samples cases or individuals that differ on some characteristic or trait" by John W. Creswell, *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research,* 4th ed. (Boston: Pearson, 2012), 207–8.
65. Triangulation in qualitative research is the use of multiple methods, for example, and "involves corroborating evidence from different sources to shed light on a theme or

perspective. When qualitative researchers locate evidence to document a code or theme in different sources of data, they are triangulating information and providing validity to their findings." John W. Creswell, *Qualitative Inquiry and Research Design,* Kindle edition, Loc 4854.

66. Carlson discusses this issue in relation to multiple studies. See Carlson, "Demystifying the Data Interview," 9.
67. Creswell identifies eight validation strategies that researchers may use to "document the 'accuracy' of their studies." Creswell, *Qualitative Inquiry and Research Design,* Loc 4843-98.
68. Lage, Losoff, and Maness, "Receptivity to Library Involvement in Scientific Data Curation."