2015

Research Data and Linked Data: A New Future for Technical Services?

Sherry L. Vellucci
University of New Hampshire - Main Campus, sherry.vellucci@unh.edu

Recommended Citation
Vellucci, Sherry L, "Research Data and Linked Data: A New Future for Technical Services?" in Rethinking Library Technical Services: Redefining Our Profession for the Future, ed. by Mary Beth Weber, 85-122 (Lanham, MD: Rowman & Littlefield, 2015).

Chapter Six

Research Data and Linked Data

A New Future for Technical Services?

Sherry Vellucci

> Data curation challenges are increasing as standards for all types of data continue to evolve; more repositories, many of them cloud-based, will emerge; librarians and other information workers will collaborate with their research communities to facilitate this process.— Association of College and Research Libraries (ACRL) Research Planning and Review Committee [1]

> [Linked Open Data] is moving to a global scale. . . . Openness, collaboration, and cooperation will bring progress and carry us forward into the world of linked open data.—Yoose and Perkins [2]

The library literature is awash with articles that discuss the future of academic libraries in terms of data.[3] Many versions of the term appear in the context of research, including research data, big data, data management, data profile, data life cycle, data curation, data services, data citation, and data literacy. What changes in research practices and emerging research needs offered librarians opportunities to engage with the research data process when their traditional support role is to build collections, provide access to the published research record, and offer assistance finding and acquiring data sets? The evolving research and publication landscape may present new and exciting opportunities for outreach and collaboration, but what does this new research data vision mean for technical services librarians?

In the context of the Semantic Web, the term *data* appears as linked data, open data, linked open data (LOD), and Web of data. For many years catalogers and metadata specialists have witnessed constant change in almost every aspect of their work—from cataloging codes, metadata standards, and record formats to digital libraries, markup languages, and the transformation of library catalogs into discovery systems. As technology and user expectations change, information organization resides in a world of perpetual beta! Just when catalog and metadata librarians are adapting to these many changes, linked data appears on the horizon with the potential to turn things upside down once again. Where did the idea of linked data come from, and will it persist as a new direction for cataloging and metadata?

A literature review confirms the importance of data in the future of academic libraries. Library participation in data curation is included in several recent "Top Trends" lists. In the June 2014 issue of the *Association of College and Research Libraries News*, the theme of the biennial article "Top Trends in Academic Libraries" is "deeper collaboration."[4] Data is cited as the top trend and includes three subcategories: (1) new initiatives and collaborative opportunities; (2) cooperative roles for researchers, repositories, and journal publishers; and (3) partnerships related to discovery and reuse of data.[5] Deeper collaboration perfectly describes the trend for librarians and researchers working more closely together. Also in 2014, the Library and Information Technology Association's (LITA) "Top Tech Trends" similarly includes research data management, big data, open data, and the Semantic Web of linked data.[6]

In their report on data curation as an emerging role for academic librarians, Walters and Skinner see the organizational structure of libraries "focusing less on the public and technical services paradigm of old (front of the house and back of the house . . .) and more on building the trio of strong infrastructures, content, and services."[7] This new type of organizational structure supports a broad view of resources and services intended to facilitate extensive collaboration for data curation and preservation services.

Research data and linked data may seem to be associated with different spheres of library activities, but both are concerned with data organization, discovery, access, and support of shared data beyond the library, and both are addressed here. This chapter examines many aspects of research data from the perspectives of researchers and librarians. It briefly examines events prior to the library's greater involvement with research data, looks at how librarians gained some fundamental knowledge and skills to assist with the tasks involved with research data curation, and discusses why researchers began to place more emphasis on data management. It then examines the stages of the data life cycle, the components of a data management plan, the purpose of application profiles, and the usefulness of standardized vocabularies and ontologies. The chapter then discusses the connection between the Semantic Web and linked data and how this integrates with research data, standards, and its future for the shared library. Metadata is a common factor among these topics. The chapter concludes with a discussion of options for librarians to expand their data expertise or retool for a new future.

WHY DATA, WHY NOW? A PERFECT STORM OR A PERFECT PARTNERSHIP?

Over the last decade, changes in many sectors of higher education converged to force academic libraries to reassess their roles and functions. The reasons behind these changes are many and complex, but the end result was an emerging focus on open access (OA) to information and sharing, repurposing, and curating research data. Rapidly advancing technology, such as expanding bandwidth for faster data transfer and lowering costs for data storage, provided the necessary infrastructure enhancements that enabled new methods for collecting and working with large research data sets. Technology was a major catalyst, but not the sole catalyst, for change. The confluence of many factors appeared to create a perfect storm, generating a data surge that challenged the ability of librarians and researchers alike to make data meaningful and useful beyond the immediate research project. We can trace the beginnings of a data paradigm shift for libraries and researchers back to several critical incidents; these include the emerging OA movement, new approaches to scholarly communication and publishing, new mandates from funding agencies requiring data management plans, and the development of institutional repositories.

Open Access and the Changing Scholarly Communication Landscape

The critical incident that drew researchers' attention to the need to manage their data more effectively came from government funding agencies and the OA movement. The combination of escalating research journal costs and shrinking library budgets meant less available access to research articles—a key driver leading to library support for OA.

The lengthy lag time between completion of research projects and publication of results posed another serious problem for both researchers and government funding agencies. The fundamental principle of scientific research is the ability to replicate the research, since scientific research is cumulative, often incorporating the findings of earlier work. It is not surprising, then, that many scientists and funding agencies found the expensive and relatively slow publishing models a hindrance to the rapid advancement of science. Unhappy with current publication practices, the National Institute of Health (NIH) proposed that papers describing research funded by their agency be deposited in an open-access database (i.e., PubMed Central).[8]

Emerging digital publishing models and ever-increasing technological capacity make it possible for published research articles to link to the supplemental data on which the study was based, thus making it available to the wider public.[9] Openly accessible research *data* makes it easier to replicate studies and repurpose data for other research projects, thus circumventing the time-consuming process of collecting data anew. But in order for data to be reused, it is necessary to provide documentation that sufficiently describes the data, ensures its quality, and makes it viable for reuse. The National Science Foundation (NSF) mandate that requires all funding proposals to provide an explicit data management plan has had a major impact on the scientific community, and it seems likely that government mandates will increase. Under the Obama administration, the government is rapidly moving forward to support OA to data. In May 2013, the President issued an Executive Order that outlined an Open Data Policy as the default model for government publishing;[10] in June 2013 the G8 leaders endorsed the five strategic principles of the Open Data Charter;[11] and in May 2014 the government followed up with the U.S. Open Data Action Plan, "which outlines new commitments as well as plans for enhancements and releases of certain data assets across the categories set forth by the [Open Data] Charter."[12] The new national and international policies underscore the importance placed on data in a global context. Some researchers see these mandates as an additional burden and have neither the time nor the expertise to manage such data plans. Librarians saw this as an opportunity to offer assistance in innovative ways that were not usually associated with librarians.

Institutional Repositories

The critical incident that prepared librarians to engage more collaboratively with researchers in managing their data was libraries' campuswide leadership role in creating institutional repositories at the beginning of the new millennium.[13] The process of designing, building, and populating an institutional repository offered countless opportunities for innovation. Professionally, technical service librarians gained or enhanced their knowledge, skills, and experience in areas such as project management and systems analysis; determining functional requirements for hardware, software, and metadata needed to support a repository; conducting research that honed survey, interview, and observation skills to discover the needs and information-seeking behaviors of different users; developing a workflow process for preparing digital objects for archival-quality preservation; working with many new metadata schemes; learning new data encoding methods such as XML, XSLT, and RDF; gaining knowledge of open-access issues; understanding intellectual property rights and copyright laws for digital objects; refining problem-solving skills; and costing out the expenses of staffing this new service and providing data storage.

Library culture began to shift as scholarly communication and technical services librarians increased their collaboration skills and raised their visibility by working more closely with faculty, information technology (IT) staff, and colleagues in other library departments. Collaboration with faculty often involved explaining open access, author rights, and the purpose of institutional repositories, as well as helping them deposit their publications. These interactions frequently enhanced librarians' knowledge of a faculty member's research and the terminology of different disciplines, while fostering trust between faculty and librarians. Collaboration with other academic units in the institution, such as the Graduate School and Honors Program, was also necessary to develop policies for depositing students' electronic theses and dissertations in the repository. Collaboration with IT staff was necessary to build the information architecture for a repository that meets the functional requirements laid out for hardware, software, and metadata. The collaborative process provided a common ground where librarians and IT technicians could learn each other's terminology, priorities, technical needs and limitations, and cultural differences. Within the library, the cross-functional nature of the work required to build, maintain, and provide a full-service institutional repository pushed the boundaries of the traditional organizational structure, and involved cataloging and metadata services, scholarly communication, and library liaisons to academic units, special collections, and archives. As frequently noted, collaboration is a critical part of research data services, and for some librarians it represents a new working model that centers on a project and team work rather than an individual narrowly focused on a single item or service.

The knowledge and skills enhanced or gained through the process of building an institutional repository laid a strong foundation for librarians to assume many of the tasks required to make data accessible for the long term. The government funding agencies' policy mandates that submission of data management plans with grant proposals compelled many researchers to seek expert assistance that allowed them to focus on their research, rather than spend valuable time learning systems for data curation. These factors make possible a synergy between librarians and researchers leading to greater collaboration on research data curation.

RESEARCH DATA

The conduct of research in all areas of study is changing and producing increasing amounts of data in digital form. In her introduction to cyberinfrastructure and data for librarians, Gold notes that "data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself."[14] Gold's statement focuses on science, but it is just as applicable to research in most other subject domains. The term *data* itself is somewhat vague in that it can apply to almost any type of evidence collected for the purpose of informing research questions. The term also applies to both types and formats of data. Types are a higher-level category including observational, experimental, derived, compiled, simulated, quantitative, and qualitative data. Formats are more specific than types and may include numeric, coded, textual, images, maps, audio files, lab notes, sensor data, survey data, field notes, samples, climate models, economic models—the list goes on. Pryor calls data "the primary building block of all information, the lowest level of abstraction in any field of knowledge, where it is identifiable as collections of numbers, characters, images or other symbols that when contextualized in a certain way, represent facts, figures or ideas as communicable information."[15] This lowest level of abstraction is often referred to as raw data. Data in this raw form frequently must be processed in some way before analysis. For example, the notes taken or recorded during a focus group interview must be transcribed and coded or categorized before they can be manipulated during analysis; or data collected by instruments in the field may need to be converted to a different file structure for analysis. Deumens et al. identified three phases, or categories, of data: *raw data* from initial observations, *intermediate data* that has been processed for analysis, and *final data* showing results of the research project.[16] Some data sets are continuously added to, or added to at intervals over time. In such cases, the data provides a "snapshot" at a specific point in time but may not be permanently fixed for years.

Today's scientific research can generate massive amounts of digital data using technologically advanced instruments that gather data at an unprecedented rate. Bartolo and Hurst-Wahl differentiate big data by the "3 Vs: Volume (how much); Velocity (how fast); and Variety (how complex)."[17] No matter the forms, types, or quantity of research data, research benefits greatly if the data is structured, described, and preserved in an accurate and systematic way that facilitates discovery, access, use, and sharing with other researchers.

Library Research Data Services

Before beginning to develop research data services, librarians should identify which services are needed by researchers on their campus and what services are already available. This will help to identify any service gaps and overlaps, avoid expensive duplication of effort, ensure that different services are based in the unit that is best equipped to offer it, and save time and money for all stakeholders.

Researchers differ in their approach to managing their research data, and their need for assistance will vary. Styles range from researchers who want the responsibility and control of the data to remain completely within the research team with no outside assistance from the library, to those who are receptive to librarians helping with the full spectrum of data services. Lage, Losoff, and Maness conducted an ethnographic study that resulted in the creation of researcher personas designed by aggregating data gathered during interviews with researchers regarding their data curation needs.[18] The various personas "represent the 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."[19] An interesting finding from this study showed that a researcher's receptivity to librarians having a role in data curation did not necessarily correlate with their need for assistance due to a lack of support from other units on campus. Instead, a greater impact on receptivity was the culture or philosophy of the researcher's particular discipline vis-à-vis sharing data and collaborative research projects.[20] These personas provide a different and creative approach to developing researcher profiles and offer useful disciplinary insight into a researcher's potential data management style that will help when deciding on and marketing new data curation services. Another approach to determining a researcher's need for assistance with a data management plan that also provided information on receptivity of librarians was undertaken by Steinhart et al., who surveyed researchers to ascertain how well prepared they were to meet the new data management plan requirements mandated by the National Science Foundation.[21] Their findings revealed that (1) researchers were unclear about the meaning of the requirements and how to implement them; (2) they were uncertain whether their data met standards in their discipline; (3) most assigned no metadata and did not know if a metadata standard existed for their discipline; (4) most were willing to share their data, although many indicated that there were circumstances under which they would not; and (5) researchers welcomed "offers of assistance—both with data management planning, and with specific components of data management NSF asks them to address in their plans."[22] Steinhart's findings are supported by a 2012 study by the Council on Library and Information Resources (CLIR), which reported that researchers queried in their study "repeatedly cited a lack of time to conduct basic organizational tasks, let alone time to research best practices or participate in training sessions."[23] These studies demonstrate a need for data management plan services, even though some researchers might be reluctant to seek assistance. Librarians should understand and respect researchers' data management styles, but they should also make sure that researchers are aware of the research data services offered by the library.

Research data services include activities that cover the entire spectrum of data curation and involve librarians across most library units. Tenopir, Birch, and Allard offer the following definition of research data services:

> Research data services are services that a library offers to researchers in relation to managing data and can include informational services (e.g., consulting with faculty, staff, or students on data management plans or metadata standards; providing reference support for finding and citing data sets; or providing web guides and finding aids for data or data sets), as well as technical services (e.g., providing technical support for data repositories, preparing data sets for a repository, deaccessioning or deselecting data sets from a repository, or creating metadata for data sets).[24]

This definition indicates the wide range of services that can be offered by the library and engages librarians who work in many library units, including collection management, acquisitions, cataloging and metadata services, library systems/IT, scholarly communications, reference, and instruction. It is clear that a majority of services listed here fall within technical services and amount to a broader role for technical service librarians than is commonly believed.

Taking a broad, institution-wide approach, Wakimoto found four key factors essential for developing research data services.[25] These included (1) collaboration, which is critical to address the full spectrum of digital curation services; (2) partnership among the three key campuswide units: the University Libraries, Research Computing, and the Office of Research; (3) sustainability of the socio-technical infrastructure for successful implementation; and (4) institutional culture as a significant factor when determining what research data services are needed and how those services might evolve in the future.[26] These key factors are echoed throughout the literature and will be the top-level metrics of success for library data curation services.

Tenopir, Birch, and Allard conducted a survey in 2012 to discover what services were being offered by academic libraries. The study found that a minority of academic libraries offered research data services, but "a quarter to a third of all academic libraries [surveyed] are planning to offer some services within the next two years."[27] This growth prediction is supported in an ACRL environmental scan[28] and by a growing body of literature[29] that indicates rapidly mounting interest and a strong future for these emerging library services.

Responses from a 2013 Association of Research Libraries (ARL) survey identified two categories of research data services: broad, long-standing library services, and services that are more specific to data management.[30] Many of the services listed in the first category are presently offered by many libraries as part of their wider service mission. Services in this broad category include the following:

- Helping researchers locate and use data sources
- Data set acquisition
- Copyright and patent advising

- Support for geospatial analysis
- Institutional repositories

Among the libraries that have implemented some form of research data services, the ARL study found the following services that were specific to data management:

- Online data management plan resources
- Data management plan training and consulting
- Digital data repository services
- Long-term data management and preservation advice
- Training/support for the software application DMPTool[31]
- Consultation on data management plan services for grant proposals
- Consultation on data management best practices (online and workshops)
- Help in identifying and applying appropriate metadata standards
- Advice on data citation, data sharing, and access
- Assistance with file organization and naming (also may be offered elsewhere on campus)
- Data storage and backup (also may be offered elsewhere on campus)
- Data archiving assistance (IR with data sets, digital repository with data sets, and data-specific repositories)[32]

In addition to these more common services, some survey participants also supported data publication, data rights management, and digital image data conversion. Not every academic library offers, or should offer, every service. The services offered by the library should fit the needs of their particular user population and the ability of the library to offer the research data services with the staff, expertise, and budget available to the library for this purpose.

User-Centered Focus for Services

Currently, large university research libraries are the most obvious institutions to offer research data services. This is supported by the survey conducted by Tenopir, Birch, and Allard, whose findings indicated that libraries in larger universities, at doctoral granting institutions, and on campuses that receive NSF funding are more likely to offer or plan to offer research data services.[33] It is understandable, then, if libraries in smaller institutions find the many articles on big data curation off-putting or not applicable to their everyday work environment. Data curation, however, is also important to research endeavors in smaller universities and liberal arts colleges across all disciplines.[34] Of course, not all data sets are "big," but that does not mean they are not important for research.[35] Carlson notes that the majority of research conducted in the science, technology, engineering, and medicine (STEM) areas is smaller research projects, with data dispersed around campus, residing on the PCs and laptops of small research groups or individual researchers.[36] Smaller, dispersed research data sets may actually present more difficulties for data curation. According to Akers, "Data from smaller studies may be captured in a variety of file formats with no standard approach to documentation, metadata or preparation for archiving or reuse, making its curation even more challenging than for big data."[37] Dispersed research data also requires librarians to make greater outreach efforts to discover and work with these researchers. Toups and Hughes stress the importance of interpersonal engagement with researchers to ascertain their data curation needs and the benefits of the smaller academic setting for face-to-face discussions.[38] On a broader scale, an environmental scan is a useful tool to identify the extent and types of research data services needed on campus and provides an opportunity for librarians to interact with researchers. A focus on user needs is the best path to a successful research data services program.

We also tend to think of data as the purview of faculty, but in fact, all types of data are increasingly collected, transformed, and used by other researchers in the academy, including students and librarians. Although graduate students are required to take courses in research methods, often there is little or no instruction focused on the process of data curation. This lack of education leaves graduate students unprepared to deal with the full life cycle of research data, which, at best, results in senior researchers using valuable time to train graduate assistants; at worst, the data may not be able to provide the necessary information to answer the research question, support the hypothesis, or be easily usable.[39] At the undergraduate level, inquiry courses are being integrated into core curricula,[40] and the growth of undergraduate research conferences demonstrates a need to grasp the importance of data management earlier in a student's academic career.[41] In this regard, data literacy is a new area for instruction—a service that should be proactively marketed and offered by librarians as part of a comprehensive plan that targets methods of instruction for specific user groups, depending on the users' existing levels of expertise.[42]

Research Data Curation

Data curation involves a set of activities that parallels the research process and ensures the current and future reliability of the research data. In its Charter and Statement of Principles, the Data Curation Centre (DCC) defines data curation as "maintaining and adding value to a trusted body of digital research data for current and future use; it encompasses the active management of data throughout the research lifecycle."[43] One of the operative terms in this definition is *trusted*, which refers to quality assurance and quality control. Procedures to safeguard quality include use of appropriate standards and best practices for data structure, organization, description, processing, and interoperability; careful and ongoing data inspection; and documenting these decisions to create trustworthy data sets that enable other researchers to replicate the study and reuse the data with confidence.

Shreeves and Cragin offer another definition of data curation as "the active and ongoing management of data through its lifecycle 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."[44] This definition identifies some tasks in the data curation lifecycle and circumscribes the length of curation to the period when the data is of interest and useful. Implicit in this statement is the notion that not all data needs to be retained indefinitely. This decision may be determined by the funding agency, the researchers' institution, or the researchers themselves, and should be reviewed periodically throughout the project.

Data curation is the area that typically engages technical services librarians. As mentioned earlier, to some extent librarians possess the fundamental knowledge and expertise to help researchers manage their data in meaningful and useful ways, especially with their metadata needs, and of course some librarians are specialists in this area. Still, a steep learning curve exists in other areas of research data curation. Data sets differ in several ways from the typical research products with which librarians commonly work. In 2007 Gold stated that "librarians are much less familiar with the data-generating research phases of the scientific research cycle than with post-research phases of reporting, communication and publication."[45] She went on to cite several challenges, including a lack of general understanding of the various methods used by e-science to collect, process, analyze, and curate research data, as well as differences of understanding between researchers and librarians of seemingly common vocabularies. Even so, recognizing several areas where they could assist in the data curation process, librarians remained keen to tackle the challenges and began to update and enhance their knowledge and skills. Three years later, Gold acknowledged that "The library profession has demonstrated significant conceptual progress in characterizing and understanding data curation both in theory and in practice."[46] One need only look at the number of webinars and workshops offered in data management, curation, and preservation to understand librarians' desire to enhance their knowledge and learn new skills to support this emerging area of their field.

The Research Data Lifecycle

The purpose of a data lifecycle model is to guide the data planning process and provide the general framework for data management and curation. To provide the needed support services for data curation, librarians must first understand the components of the research process and the phases of the research data life cycle. The best practice is to develop the data lifecycle model in conjunction with each stage of the research plan (often called the *research* lifecycle model). A research data life cycle groups related activities together into stages that are often sequential and dependent on the previous stage. For example, the planning stage involves deciding on the type of data that is needed to answer the research questions and determines how the data will be gathered and stored. The next stage involves collecting the data, but this is dependent on the data specifications determined in the planning stage. Some stages in the data life cycle may be iterative—for example, after data is collected and examined, new or additional data might be required, and the collection stage begins again. The groups of activities continue throughout the data life cycle and on into the archiving and preservation stages after the research is completed.

The research data life cycle is represented as a graphic conceptual model at a high level of granularity. Each subject domain and research project tends to develop its own data life-cycle model that fits its particular needs. For this reason, no model is one-size-fits-all, but there are life-cycle stages that are common to most research and can be adapted to a particular research project. The Committee on Earth Observational Satellites maintains an ongoing compilation of data life-cycle models that currently consists of fifty-five models across various disciplines.[47] Some of the models included are quite simple, while others are much more complex.

Figure 6.1 is an example of a simple life-cycle model. Adapted from a model developed by the United State Geologic Survey (USGS),[48] it uses a linear matrix structure expanded to eight sequential primary life-cycle stages. The matrix below the primary-stage arrows indicates processes that are non-sequential and ongoing throughout all phases of the research project.

The *research* lifecycle outlines all stages of activity for the entire research project, while the *data* lifecycle covers the stages where some type of activity involving data occurs. In reality, the data lifecycle is embedded in the research lifecycle. For this reason, many models combine the research and data

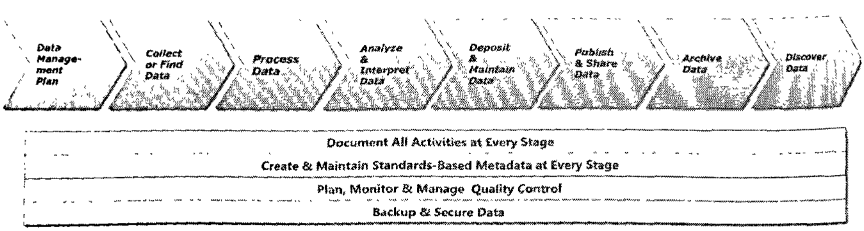

Figure 6.1. Linear Matrix Life Cycle Model. ***Adapted from the USGS Linear Data Lifecycle Model***

lifecycles into one, which contextualizes the data within the overall research project. Figure 6.2 shows the research lifecycle model from the University of Virginia and is an excellent example of a contextual model that includes both the research life cycle and the data management lifecycle.[49]

The Virginia design expands on a linear model, beginning with the first stage in the *research* lifecycle—planning and writing the research proposal—and continues through the end of the research project and the data archiving process. The Virginia model also incorporates features of a nonlinear or cyclical model that includes recursive stages for additional data discovery, collection, analysis, repurposing, and reuse. This contextual model includes data preplanning as part of the research proposal writing phase. During proposal writing, librarians can assist researchers with discovery of existing data sets and begin to gather information that identifies the types of data and data events that will occur during the research project. This stage also provides an opportunity for librarians to gain knowledge of the nature of the data being collected and to help structure the data model for optimum access, use, archiving, and sharing. In the Virginia model, the detailed data management plan is developed in the "Project Start Up" stage, which begins after the project is funded. These first two stages are sometimes considered conceptual or administrative in nature because much of the activity entails planning and testing research tools and processes.

The more dynamic data lifecycle phases begin with data collection and run through data analysis and data sharing; recursive activities with data can occur anywhere during these stages. The analysis phase involves processing data to enable analysis and interpretation that facilitates sense making for researchers to grasp the implications of the data as they relate to the research

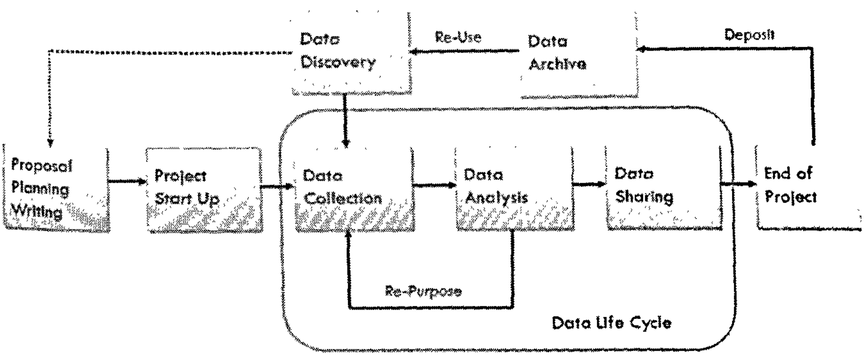

Figure 6.2. The Research Life Cycle. ***Figure reproduced with permission from the University of Virginia Library, Research Data Services. © 2014 by the Rector and Visitors of the University of Virginia***

questions. There is always the possibility that data-gathering instruments may need recalibration, or the data may need to be reprocessed, updated, and cleaned (i.e., duplicates removed, checked for missing data or other errors). Data sets might also be merged at some point to integrate new data or conduct a meta-analysis. Data manipulation makes identification of different versions of the data a critical part of ongoing data maintenance throughout the project. The data is then prepared for sharing. This could involve structural or format conversion to a required file format standard or, if not already completed, stripping the data of any information that could identify participants, if the research involved human subjects, or other sensitive data. When the data is configured for sharing and the research results written up for publication or other forms of dissemination, the data set should be fixed at that point in time to represent a snapshot of the data used to support the results of the published article.[50] The data snapshot is then archived, but that does not necessarily mean that data collection is completed. Many longitudinal studies will continue to generate data and produce project reports over time. A new fixed data snapshot should document the data used in each report or published article.

The end of the research data lifecycle deals with postresearch data activities that include data archiving, discovery, and preservation. Open access databases and some subscription journals link the published research article directly to the supporting data, rather than making the reader search elsewhere to locate and reuse the data from the study. At this point, the data can either be discarded from short-term storage or archived and preserved in long-term storage for the length of time determined in the data management planning document.

It should be noted that not all researchers are eager or able to share their data. In addition to issues of confidentiality and privacy, other reasons for withholding data publication include research that results in proprietary products, time embargos imposed by funding sources or industries, or a desire to publish the research findings in articles prior to releasing the data.[51] Cragin et al. conducted in-depth interviews with researchers to examine data sharing in more detail and found that research discipline was not the predominant factor in willingness to share research data.[52] Researchers were most willing to share their data with other colleagues in the same field. Also, "willingness to make data available increased as data were cleaned, processed, refined and analyzed in the course of research."[53] Several researchers were concerned due to personal experience, where their data had been misused, misinterpreted, and cherry-picked by industry to support efficacy claims about their products.

Another problem of concern to researchers is the lack of attribution or inaccurate data citation by other researchers who used their data.[54] International groups are working to make data identification and citation easier.

DataCite is dedicated to helping researchers discover, access, and reuse data.[55] The DataCite membership took the lead in developing the "Joint Declaration of Data Citation Principles"[56] and currently focuses on working with the publishing community to assign digital object identifiers (DOIs) to data sets to provide unique identifiers. While this cannot negate ethical problems, it certainly is a step in the right direction for proper citation. DataCite has also created a metadata scheme and best practices to facilitate consistent and accurate citation of data sets.[57]

When first implementing new research data services, a researcher's lack of awareness of the data services offered by the library may be a barrier to collaboration at the outset of the research planning stage. Librarians are often called upon for help in the middle of a project lifecycle, after data is collected, or in other cases when the project is finished and the researchers are seeking a place to archive their data. Librarians can enter the data lifecycle at any point to help with data curation, but the further into the research lifecycle, the more difficult it may become to apply best practices to the data curation process. The activities for each stage of the research lifecycle model provide context for the stages of the data lifecycle model, which in turn provides a framework for the data management plan.

Data Management Planning

A data management plan provides the written documentation that describes the data and the data events needed to support each stage of a research project's life-cycle model and identifies the persons responsible for each activity. Just as development of the data life-cycle model overlaps with development of the research life-cycle model, an outline of the data management plan will begin to take shape as decisions are made regarding the data life-cycle model.

It may seem counterintuitive to say that priorities for crafting the data management plan should focus on the researcher, not the data, but adopting a user focus will ensure that the data is fit for purpose. When the focus is solely on the data, the end result may be an elegant data management system that cannot be used by others or is not interoperable with other systems used by the particular discipline for discovery, access, storage, retrieval, and archiving. The best way to ensure that the user remains central to the plan is to work as a data management team, with members from the research project, library, research computing center, and other campus units with the required expertise.

Within a broad context, a data management plan can be viewed as consisting of three primary stages: the active research data phase (data planning prior to and during the research project), the end-of-project phase (review data documentation and data sharing plans, check adherence to ethical and legal policies), and the final postresearch archiving phase (review retention specifications, data growth, access, security plans, and preservation plan). Thinking in this broad context at the inception of the plan will enable decisions made for each component of the data management plan to result in the most effective use, reuse, and storage of the data.

Data management plans may take many forms and include different component parts. Data management plan checklists, which outline details of the content to include in the plan, are readily available.[58] While these checklists vary, some common components include an introduction, data description, documentation, metadata, intellectual property rights, backup and storage, security, dissemination and sharing, maintenance, archiving and preservation, data attribution, and plan administration. Figure 6.3 offers basic questions to consider for each category when developing a data management plan.

Most data management plans found in the literature are written for researchers, not librarians, and it may be unclear exactly what role the librarian has or where a librarian fits into the overall process of developing a data management plan. Ultimately, the researchers will make this decision—hopefully, with knowledge of the data services available to them. Generally, librarians serve as consultants in their areas of expertise, extending their knowledge to the research process and data collected in a given discipline to perform successfully as a data management plan consultant.

A good way to start a collaborative data management plan partnership is for librarians to reach out to researchers to create data curation profiles.[59] Developed by Purdue University and the University of Illinois, the data curation profile is an instrument that provides a structured method to outline and document information about a researcher's data needs throughout the multiple stages of the research process in a particular subject area.[60] The process for developing a data curation profile starts with the librarian conducting an in-depth interview with the researcher to identify requirements related to data curation in their particular area of research. That information is then used to create a profile in a concise, structured document that is suitable for sharing.[61] The profile information can then be incorporated into a data management plan. Because data curation profiles register the details of the research process in a specific subject field, they can be adapted to serve as a basis for developing future data curation plans for other research conducted in that subject domain. Purdue maintains a data curation profiles directory containing completed subject profiles that have been reviewed and deposited for sharing, reuse, and preservation.[62] Not only do data curation profiles provide in-depth details of a particular research process, they help break down cultural barriers by demonstrating the librarian's desire and ability to understand the researcher's work and data needs, which helps to develop a more trusting, collaborative relationship.

Data Management Planning Questions

Introduction: What is the context for the research and the data management plan? Who is the funder and what is the project ID? What internal and external policies apply to the data (funder, institution, researcher, journal)? Who is the contact person for the project? The data?

Data description: What type and format of data are needed to inform the research questions? Is there existing data? How will the data be collected (methods, equipment, software)? What will be the growth rate and will it change frequently? What conventions will be used for file organization and directory and file naming? Are tools or software needed to create/process/visualize the data?

Documentation: What documentation is needed to replicate the project and reuse the data (data collection procedures, codebooks, data dictionaries, data definitions, data processing and analysis methods, reports, publications, website)?

Metadata: What metadata standard, controlled vocabularies or ontologies will be used to describe the data? Do discipline specific standards exist? How will the metadata be created or captured? How will the different versions of data be tracked? Will technical, administrative and rights management metadata be used to facilitate discovery, retrieval, interoperability and sharing of the data?

Intellectual property rights: Who will hold the intellectual property rights to the data, and how will IP be protected if necessary. Will a Creative Commons license be used? Are there any copyright constraints (e.g., copyrighted data collection instruments) should be noted.

Backup and Storage: How will the data be backed-up and how often? What volume of storage is needed? What media/formats will be used? Where will the data be stored during the project? How will the data be stored, maintained, archived and preserved for the long term? What version(s) of the data will be retained and for how long, or will the data be discarded at the end of the project? What repositories are available for data storage and deposit? Are there subject specific repositories?

Security: What type of security is there to safeguard the data? Are there confidentiality, privacy, commercialization, high-security or ethical restrictions? How will data access be managed during the project?

Dissemination and sharing: Who is the audience? How will potential users discover and retrieve the data? Where and when will a report/article be publish? Will the data be shared? Will it be linked to the article? Is there an embargo period on access to the data and if so, how long? What tools/software are needed to access, visualize and retrieve the data? Will data access be mediated?

Maintenance, archiving and preservation: What data should be retained, shared and/or preserved? What are the foreseeable research uses for the data? Will the final format of the data files be in a sustainable format for long term access? Will supporting documentation be archived with the data? Who maintains the data after the project is completed? Are there subject archives suitable for the data? Are there costs for long-term storage?

Data attribution: How should the data be cited? Will the publisher assign a DOI? Where will the desired citation form be published?

Plan administration: Who is responsible for each activity or event in the data process; and how much will each activity or event cost? Who will assure that the Data Management Plan is followed? When will the plan be reviewed or revised?

Figure 6.3. Data Management Planning Questions

Metadata for Research Data

"Metadata is structured data that describes a resource [i.e., research data], identifies its relationships to other resources and facilitates the discovery, management and use of a resource."[63] It provides the contextual basis to identify, authenticate, describe, locate, and manage research data and related resources in a precise and consistent way. Metadata can help researchers find their own data more efficiently by bringing consistency to all documents and data in a research project. When subject-specific metadata are applied, they extend consistency to other research projects in that same discipline. Using a standard metadata scheme to describe a data set facilitates interoperability and makes integration of data sets much easier. With the new cultural shift toward research transparency and data sharing, many journal publishers and subject repositories now designate specific metadata requirements for data deposit.

In the context of a data management plan, a strong case can be made for assigning metadata to research data, but many researchers are confused by the term *metadata* and do not understand the important functions that metadata serve. Based on a study conducted at Purdue University, Carlson notes that researchers "often do not have a sense of what metadata should be applied to their data set to enable it to be discovered, understood, administered or used by others."[64] Discipline-specific metadata schemes can be extremely complex, making them difficult and time consuming to apply. In response to this problem, many schemes have identified a *core set* of elements that create a basic level of description; examples include Dublin Core, Darwin Core, and TEI Lite. Researchers can also select essential elements from a specific scheme to create their own core metadata, but this should always be documented for future understanding and use. Metadata librarians can provide a wealth of knowledge and expertise to assist with these activities, but it is important when developing library data services that outreach efforts to researchers are clear, precise, and jargon free.

Types and Functions of Metadata

Many metadata standards are developed and maintained by discipline communities to support the specific needs of that group. Each community tends to define different types of metadata grouped by the functions they perform. Greenberg defines the general functions of research metadata as discover, manage, control rights, identify versions, certify authenticity, indicate status, situate geospatially and temporally, and describe processes and architectural structure.[65] Librarians typically think in terms of the more generic FRBR (Functional Requirements for Bibliographic Records) user tasks—find, identify, select, and obtain—and categorize metadata functions as descriptive, administrative, and technical, containing subcategories such as intellectual property rights and usage, events, preservation, and structure. There is not universal agreement on the definitions for these categories, and subcategories are often considered primary categories in their own right. These three broad categories are not mutually exclusive. Metadata elements can overlap boundaries, complicating any categorization of elements. For example, is the time period of data collection a descriptive element, an event, or both? Regardless of how they are grouped, the important consideration is that the functional requirements are appropriate to the specific research project and data sets.

Descriptive research metadata performs functions similar to metadata used in library catalogs but uses many different elements and standards. In the broad view, research metadata is used to enable researchers to *find* the data sets based on their search terms, *identify* which data sets are appropriate for their research needs, *select* the best data set for their purposes, and *obtain* the selected data set. Descriptive metadata for research data sets, however, will vary depending on the subject domain and type of data and will contain data-specific elements such as data identifiers; methods of data collection and source of data; variable names, labels, and groups; sampling procedures; greater specificity/technical subjects; geospatial and temporal information; and data citation form. Descriptive metadata equates to Greenberg's functions of discover, identify versions, situate geospatially and temporally, and indicate status.

Administrative metadata provides information that is needed to manage and maintain data sets. This is where much of the data life-cycle information is recorded. These metadata elements might provide information about the data management plan, including names and contact information of persons responsible for various data-related tasks; file sizes and naming conventions; data processing methods; equipment and software used to process and analyze data; maintenance and update frequency; status of the data (raw, processed, final); tools necessary to view the data; and data storage information. Administrative metadata also includes rights information about ownership, intellectual property, usage data (including restrictions on use, reuse, and sharing); and event data (including information about data capture, integration with other data, observational occurrences, information about data processing and analysis, and data retention or disposal information). Administrative metadata equates to Greenberg's functions of manage, control rights, certify authenticity, and describe processes.

Technical metadata includes information on the metadata scheme structure used; the technical specifications for format types, encoding, viewing, processing, storing, and exchanging data sets; data validation methods; and information needed to migrate and curate the data for long-term preservation as software and hardware change over time. Technical metadata may also include structures used to link the data set to related objects such as journal articles, different versions of the data, and additional forms of documentation related to the data set.[66] Technical metadata equates to Greenberg's functions to manage, certify authenticity, and architectural structure. Consistent data values (i.e., standards) are critical to all categories of metadata to ensure the most useful data organization and description, provide relevant data retrieval and integration, and facilitate preservation, sharing, and reuse.

Standards

The Environmental Protection Agency offers this definition for data standards: "Data standards are agreed-upon definitions, formats, and procedural rules for commonly used data sets that are needed to reduce the complexity of data manipulation and to make the exchange and integration of data more efficient."[67] Standards are an essential part of creating high-quality, trustworthy metadata, for without standards automatic metadata creation, discovery, interoperability, data sharing, and long-term preservation would be a considerable challenge. Key to the success of data curation, standards provide tested and trusted models that define best practices. In the metadata environment, several types of standards are used, including those for structured metadata schemes, content, vocabularies, and syntax. Some standards (i.e., standard metadata schemes) may incorporate other standards, such as directions for how to enter data content and specifications for what vocabularies to use for data values and what syntax scheme to use.

Structured Metadata Schemes

Standardized metadata schemes define the number and types of elements in the metadata structure and range from a simple general scheme such as the Dublin Core Metadata Element Set (DCMES), with only fifteen elements designed to describe all types of resources at a high level of granularity, to subject-specific metadata schemes with dozens of elements that provide great depth of detail, such as the ISO 19115-1:2014 Geographic Metadata scheme.[68] The DCMES is the most commonly used international metadata standard, largely due to its high level of description and simplicity, which enable it to function as a switching language to facilitate interoperability and easy metadata harvesting. When metadata schemes other than DCMES are used to describe a data set and are then mapped to the DCMES elements, it expands the ability to share research data exponentially.

Subject-specific metadata schemes proliferate across the wide spectrum of research disciplines and it would be impossible to cite them all here. Determining which metadata standard is most appropriate for a given data set is a complex and, as noted, confusing decision. The Data Documentation Initiative (DDI) metadata scheme is used widely for social science data sets, for it uses a hierarchical structure that describes the research project, the data sets used in the project, and the variables used in the data sets.[69] The ISO 19115-1:2014 Geographic Metadata scheme is used by many disciplines, but a majority of subject-specific communities have devised their own metadata. To alleviate the problem of finding the right metadata, the Data Curation Centre (DCC) maintains a discipline-based list that includes not only information about a metadata standard but also application profiles, tools to implement the standards, and use cases of data repositories that currently implement them.[70] The five broad DCC disciplines include:

1. Biology (10 schemes, 13 extensions, 12 tools, and 30 use cases)
2. Earth Science (9 schemes, 16 extensions, 14 tools, and 27 use cases)
3. Physical Science (7 schemes, 6 extensions, 13 tools, and 19 uses cases)
4. Social Science and Humanities (5 schemes, 3 extensions, 6 tools, and 12 use cases)
5. General Research Data (8 schemes, 4 extensions, 5 tools, and 8 use cases)

The DCC listing has a strong European focus and is by no means comprehensive. Other subject-specific metadata lists include the JISC Digital Media Guide[71] and the Society of American Archivists Metadata Directory.[72] The most comprehensive inventory of metadata standards for the cultural heritage domain is Riley's visual mapping of the metadata landscape, which includes over one hundred standards for metadata schemes, content standards, controlled vocabularies, and syntax, structural, and markup standards.[73] Large government agencies such as the National Aeronautics and Space Administration (NASA),[74] National Oceanic and Atmospheric Association (NOAA),[75] and the EPA[76] have research data and metadata standards for their agencies' projects to provide consistent data description for data sharing and integrating with other data sets. Finally, many research university libraries regularly include lists of metadata in their guidelines for data management plans. Guidelines that were created using the Springshare LibGuides tool are open access and can be searched on the LibGuides Community website.[77]

Content Standards

Content standards are the formal rules that provide general guidelines or detailed requirements for each metadata element. This type of standard acts as an input guide for content for metadata elements that are not automatically populated or when no controlled vocabulary or syntax applies.[78] A content standard might explain how to enter a person's name, when to use capitalization, or how to format a date. Content standards are not new to librarians, as all cataloging rules up to Resource Description and Access control metadata content. Examples of other formal content standards include Cataloging Cultural Standards (CCO), Describing Archives: A Content Standard (DACS), the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata (FGDC CSDGM), and Access to Biological Collections Data (ABCD). Informal content standards are the best practices defined when a new subject-specific metadata is created and should be recorded in an application profile.

Content standards also specify what information to use for the values of specific fields. For example, a content standard might stipulate what label is assigned to a particular metadata element; indicate whether a specific element is mandatory, recommended, or optional; or identify a specific thesaurus to be used for the subject or spatial elements.

Vocabularies, Taxonomies, Ontologies

Controlled vocabularies are another type of standard that is used to bring accuracy, clarity, and consistency to metadata. In its broadest sense, a controlled vocabulary is any standardized list of terms, or set of words, selected by a community to describe and bring meaning to a data set or a group of data sets. Some vocabularies control for synonyms, disambiguate similar objects, or are used for indexing and provide a semantic relationship structure. Vocabulary control greatly improves precision and relevance in retrieval by identifying and collocating equivalent and related objects. Research data semantic vocabularies are frequently designed to be specific to the terminology used by the subject community.

The continuum of vocabulary control moves from the simple list, subject heading lists, and authority files to the more complex thesaurus, taxonomy, and ontology structures. A list is a simple group of terms for a category of things. The terms may be coded data to use as surrogates for such things as countries, languages, variables, and roles. They are widely used in dropdown menus to limit choice as a feature of a metadata input template and are especially useful in the research metadata context to assist data entry and ensure consistency.

Both authority files and subject heading lists are an integral part of technical services workflow, needing no further explanation here. Many general and subject-specific authority files are used to describe research data, and domain-specific metadata best practices will identify these.

Over the years, the Library of Congress Subject Headings (LCSH) have moved in the direction of a thesaurus by using standard relationship indicators, but LCSH is excluded from the thesaurus category because it does not maintain a hierarchical structure. Subject heading lists describe a wide range of topics and do not contain the level of subject depth that is needed for a very specific subject discipline, but depending on the type of research, the LCSH may be adequate for some data sets.

Increasing in complexity, taxonomies, also called classification schemes, are collections of controlled vocabulary terms, or a classification code, organized into a hierarchical structure primarily in an online environment. Each tier of the hierarchy inherits the attributes of the tier immediately above and becomes more specific as you move down the hierarchy. Taxonomies are limited to broader/narrower relationships with no particular entry term, since any term may be part of multiple hierarchies.[79] This allows the context of the vocabulary to change depending on the preferred term at the top of the taxonomy and makes them useful for searching in the online research data environment. The term *taxonomy* has become a catchall used by many in the online environment to refer to any type of vocabulary, causing confusion by ignoring distinctions among the different types of subject standards.

A thesaurus represents the vocabulary of a specific subject; it adds layers of complexity to a taxonomy and differs in a number of ways. A thesaurus is arranged in a known order and strict hierarchical structure so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators (BT, NT, RT, etc.).[80] Not all terms in thesauri are preferred terms, so in addition to the broader/narrower relationships used in taxonomies, a thesaurus includes *Use For* (UF) and *Use* to identify the preferred term and refer from the nonpreferred term, thus controlling for synonyms. Thesauri also include relationship terms (RT) for subjects that are related to the preferred term but not included in same hierarchy. Context is provided in thesauri in the form of scope notes (SN). Thesauri are among the most heavily used controlled vocabularies for data and are available for most subjects ranging from the *NASA Thesaurus*[81] to the *Art and Architecture Thesaurus*.[82]

The final type of controlled vocabulary on the continuum is the ontology. Ontologies define the concepts and semantic relationships in a specific knowledge domain and are presented as a formal model. Ontologies represent the most advanced point on the continuum, in the sense that taxonomies and thesauri are fixed vocabulary languages for subject description, while ontologies have open vocabularies.[83] In a closed vocabulary language, the number of concepts is open, but the language used to describe them is considered closed if the allowable relationships are fixed and limited. Since taxonomies allow only one relationship in the hierarchy (BT/NT) and thesauri use a limited number of relationship types (BT, NT, RT, UF/Use, and SN), both are closed vocabularies. Ontologies, on the other hand, allow the creator of the subject description language to define many more semantic relationships—for example, hasColor, isSister, hasDisease—making it more flexible and robust.[84]

As with taxonomies, people use the word *ontology* to mean any type of controlled vocabulary, adding to existing terminological confusion. Ontologies, like taxonomies and thesauri, are developed for a specific domain, making them an important vocabulary resource for research data sets. The original Greek term meant "the science of being" and was used by a branch of metaphysics that attempted "to organize the universe and its components into a scheme with explicit formulation of their possible relations."[85] With the advancement of technology in the late twentieth century, the term was co-opted by the information architecture and science domain and repurposed for use in information retrieval.[86] As an open vocabulary with many allowable semantic relationships, ontologies retain hierarchies but are increasingly structured as a Web network and are largely developed outside the immediate library community. Because of this, the language used in defining the components and structure of ontologies may sometimes be a bit unfamiliar to librarians, but most of the concepts are not. Gruber offered the most well-known and succinct definition of ontology: "an explicit formal specification of a shared conceptualization."[87] Vanopstal et al. offer an excellent explanation of Gruber's definition by parsing the component parts:

> Firstly, "explicit" means that the concepts included in the ontology are clearly defined, as are the constraints on their use. "Formal" refers to the language of the ontology. A formal language is computer-readable: the computer "understands" the relationships—also called "formal semantics"—within the ontology. This way, they can be used to support computer applications. . . . The last components of the definition, *shared* and *conceptualization*, imply that this abstract model of phenomena in the world has been agreed upon by a group of users or experts.[88]

Thus an ontology has clearly defined terms, which are vetted by experts in the subject domain, and tells how they can be used; however, the term *formal specification*, or *formal language*, might prove confusing. In this context, it does not refer to the language of subject terms; it means a type of structural language, such as the Resource Description Framework (RDF).

Ontologies consist of concepts, properties and instances. Concepts are the primary element and are grouped into classes. An instance (also called an individual) is a specific member of a class, that is, the real-world object referred to by the concept. Properties, or attributes, define the relationships between classes and instances. Both classes and properties may have subclasses and subproperties (a parent-child relationship), both of which inherit the characteristics of the parent class or property.

When the terms and relationships are clearly defined and a formal structural framework such as RDF is implemented, computers can use the hierarchical class/subclass structure to make inferences about like things. For example, if the subclass *cuckoo clock* is a type of class *clock*, and class *clock* is a type of *timepiece*, then a *cuckoo clock* must be a type of *timepiece*. When encoded in RDF-based ontology language, such as the Resource Description Framework Schema (RDFS) or the Web Ontology Language (OWL), the ontology becomes a computer-actionable model that enables logical inferencing.[89] Tools exist to help create ontologies. *Protégé* (http://protege.stanford.edu/) is one of the more popular applications for this purpose.

Syntax Standards

Syntax standards are the final type of standards discussed for metadata in this chapter. They provide the encoding/packaging to make the metadata machine readable and enable data to be integrated and exchanged with different systems. Librarians have worked with electronic syntax standards for many decades in the form of the MARC record format. In recent years, with the advent of the web, new markup languages are used.

The mother tongue of markup is the Standard Generalized Markup Language (SGML), from which other markup languages are derived. The most commonly used encoding standard today is the eXtensible Markup language (XML), which is designed to *describe* data and focuses on what the data is, as opposed to HTML, which was designed to *format and display* data on a web page. By removing the display function, XML becomes a more flexible and powerful syntax for exchanging data—the primary function needed for linking and transferring data on the web. Since no tags are predefined, an XML schema can be created for a general or specific subject domain, making it adaptable for metadata and research data sets. XML is used widely by metadata schemes to provide the encoding structure. Other types of XML applications extend the functionality to enable searching. One important feature of XML is the ability to use namespaces, a type of uniform resource identifier (URI) that links directly to the vocabulary term. Rather than entering the literal name for a concept like *Brooklyn Bridge,* a URI can be entered that links to the specific term in the Library of Congress Subject Authority File (http://id.loc.gov/authorities/subjects/sh85017204).

The different types of standards used for metadata are the critical underpinnings for systems that describe, maintain, process, discover, integrate, exchange, share, reuse, and preserve research data. A vital accompaniment to a metadata scheme is documentation of the standards employed and the best practices for using the scheme. This type of documentation is especially important when project personnel change, or when altering an existing metadata scheme.

Application Profiles

Sometimes no metadata scheme exists that can meet all the needs of a particular research project. To avoid a proliferation of schemes and duplication of work, one should check that no subject scheme already exists that could be adequate for the specific community or local project needs. Modifying an existing scheme could be a useful option, but the modified scheme should follow accepted standards to allow metadata and data sets to be interoperable with data in other repositories. The modified scheme should be fully documented so others will be able to find the data and reuse the newly modified metadata scheme for similar projects.

If only a minimum number of elements from one metadata scheme are used, this is referred to as a *core* scheme. For example, if only ten elements of the ISO 19115-1:2014 metadata standard are used, the modified scheme might be called a Geographic Core Element Set. A modified scheme is a new *application* of the existing scheme that should document, or *profile*, the way it will be used or extended. The final result is creation of an application profile (AP). APs consist of data elements drawn from one or more metadata schemes and customized for a particular local application.

Many application profiles exist, as well as AP data models for creating new ones. The Dublin Core Metadata Initiative (DCMI) group created the Dublin Core Application Profile (DCAP) model to assist the AP development process.[90] "The DCAP defines metadata records which meet specific application needs while providing semantic interoperability with other applications on the basis of globally defined vocabularies and models."[91]

The first step in the AP development process is to determine the functional requirements, or what you want the metadata to accomplish—for example, interoperate with other schemes, assist in identifying intellectual property rights, provide information for preservation migration, and so on. The next step is to create a domain model that characterizes the things that exist in the subject community's universe that the metadata will describe, such as the concepts, people, events, and relationships that exist between them—for example, temperature data recorded by person X using instrument Y. Following the domain model, identify the list of metadata elements needed for the metadata scheme (for the data concept alone, this domain needs elements for data set identifier, name, type, date, version—the list goes on for each element needed for persons and instrument type). The next step is to define the specific elements and provide guidelines for how they are used. This data set description is usually formatted as a table and provides, at the minimum, the element name, label, status (mandatory, recommended, optional, etc.), repeatability, type, and source of data value, standards used for vocabularies and syntax for each element. Figure 6.4 shows the more complex Singapore Framework, which provides a three-tiered contextual model of the DCAP. The top tier diagrams the *application profile* layer; the middle tier diagrams the *domain standards* used; and the bottom tier shows the *foundation standards*, including the metadata syntax and structural standards. The things documented will vary depending on the environment in which the metadata will operate. If the environment is the Semantic Web and linked data, the content categories will include *namespace* information as the source for a particular element.

The Semantic Web and Linked Data for Libraries and Research

Interest in ontology development has surged due to the convergence of three factors: (1) research data migration to the digital environment, (2) the web providing the platform for linking data, and (3) the Semantic Web gaining a foothold among researchers and librarians. As noted above, ontologies are constructed to function in the web environment and rely on sophisticated linking architecture for their structure and use. To understand better the functional application of ontologies, therefore, we must be familiar with the Semantic Web and linked data.

Miller describes the current web environment as a web of linked *documents* with unstructured data that relies on manually created hyperlinks that do not specify the nature of relationships among records; searches that match keywords in documents; and relevance-ranking algorithms. This creates a system suitable for humans but not terribly precise and of little use to machines.[92] In contrast, he describes the Semantic Web as a distributed web of *linked data* that connects structured, semantically meaningful metadata using semantically meaningful links, making parts of the web similar to a database with database functionality. The use of standardized vocabularies along with this linking data architecture makes the Semantic Web system suitable for

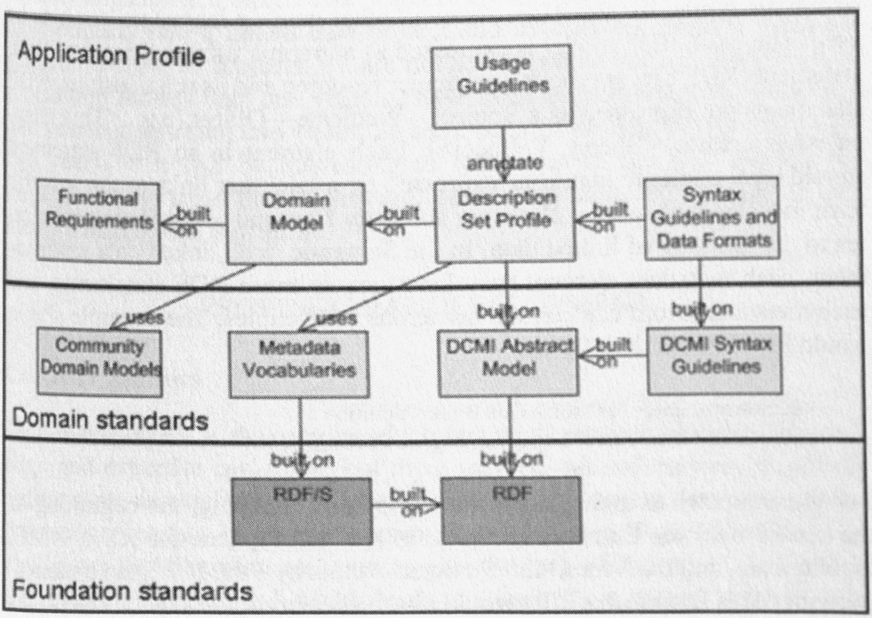

Figure 6.4. The Singapore Framework. ***Reproduction of the DCMI Singapore Framework. http://dublincore.org/documents/singapore-framework***

machine queries that will result in greater retrieval of more relevant search results and the ability to integrate data sets more easily, improving access for humans.[93] Openness is the bedrock for both the Semantic Web and linked data, for if data remains encapsulated in closed databases or library catalogs, where it is accessible only by searching each database separately, the Semantic Web and the linked data structure will be limited in its retrieval capacity. The cultural shift to open research data reinforces the importance of both the Semantic Web and linked open data capabilities.

What Miller describes as the Semantic Web is the architecture for *linked data*. In essence, linked data represents metadata *records* deconstructed into individual components of *data* (i.e., elements), rendering the data within the record structure independently retrievable. To accomplish this and maintain the context provided by the full metadata record, it is critical to use controlled vocabularies for consistency and a structure that provides for inclusiveness of a metadata record, while separately linking to the individual elements. The Resource Description Framework (RDF) data model meets this need.

The extensive literature on the RDF data model is coupled with literature on linked data. As the MARC communication format provides the structure for library catalog records, RDF is the communication architecture for linked data, and as MARC tags are the syntax for catalog records, XML is the normative syntax for the RDF, although other markup languages can also be used. The basic RDF model is expressed as a graphic using a three-part RDF statement (RDF triples) about a particular resource and its relationships. RDF statements are expressed as a Subject—Predicate—Object (e.g., <This chapter><has creator><Sherry Vellucci>). Each element in an RDF statement should have a unique identifier expressed by a URI that links to the specific term in a standard source, elevating authority files and controlled vocabularies to the gateway of linked data. In the Semantic Web linked data environment, each metadata element may have one or more RDF statements, and each metadata record can contain numerous RDF triples. The example above would be expressed in RDF as:

```
<dc:title xml:lang="en" rdf:>This Book</dc:title>
<dc:creator rdf:resource=http://viaf.org/viaf/106125310"/>
```

For the computer to understand what the syntax is saying, the beginning of the record declares that this record uses the namespaces for RDF, XML, Dublin Core, and the Virtual International Authority File. If we had a namespace for "This Book," that URI would also be declared.

While research data curation is becoming an established service, libraries have been slower to transform catalog records into linked data. Although linked data may be easier for catalogers to grasp because it deals with the elements of the familiar bibliographic record, conversion to linked data would involve enormous numbers of catalog records and a better understanding of RDF and XML. Recently, however, linked data appears to be gaining support from the library community with growing evidence that OCLC and the Library of Congress (LC) are exploring the possibilities for making linked library data accessible to a global user population.

LC has implemented a Linked Data Service,[94] which offers an impressive array of authority file, general and specific controlled vocabularies, classification, MARC languages, code lists, type lists, and so on converted into URIs and many different formats, including RDF and XML. This extensive project provides preformatted values for many components of linked data triples that will be useful for research data sets and lowering the bar for libraries to enter the Semantic Web community. For examples of linked data in context, Southwick and Lampert's presentation on RDF and linked data offers an excellent overview of its use for a library digital collection, accompanied by many useful examples.[95]

As of this writing, OCLC Research is exploring several different aspects of linked data, including an extensive survey to discover who is currently using linked data; which institutions are consuming or publishing linked data and why; and descriptions of technical details, production examples, and implementer advice. The results show a small group of respondents have actually implemented linked data projects. Of the 122 respondents, ninety-six have implemented a linked data project and seventy-six projects were described. "Of the 76 projects: 27 are not yet in production; 13 have been in production for less than one year; 12 have been in production for more than one year but less than two years; 24 have been in production for more than two years."[96] It will be interesting to follow up to see how quickly the implementation trend moves forward, particularly after OCLC fully implements a linked data WorldCat.

PROFESSIONAL DEVELOPMENT

Cultural Barriers

Data management is the area in which technical services librarians' knowledge and expertise can shine, but there may be cultural barriers to address before some researchers are willing to change their current data management practices to enable librarians to provide this service. Pryor identified two challenges that librarians must overcome: the lack of "parity of esteem with the research community," which leads to a lack of credibility; and a need to "persuade and demonstrate that they have a material contribution to make, one that is likely to be of tangible benefit to researchers and the research programme."[97] To collaborate effectively, it is incumbent upon librarians to understand the nature of the research process, the types of data, and data collection methods in a researcher's subject field to establish credibility and trust. Liaison librarians have a depth of knowledge in specific disciplines and subjects and could serve as internal consultants to help librarian colleagues expand their knowledge base. Preparing for and conducting an interview with a researcher to develop a data curation profile is one of the best ways to learn about the research process in a specific subject domain.

Resources for Professional Development

Prior to focusing on the details of data management planning and curation, librarians should begin by familiarizing themselves with the policies and procedures of the local institution; this information is usually maintained by the Office of Research. Websites of government funding agencies provide guidelines to meet their data management plan requirements, and these should always be followed when submitting a grant proposal. Page length limitations are often set for the data management plan portion of a funding proposal, which can restrict the amount of detail included, but that does not preclude adding greater depth of detail to the project's operational data management plan. Librarians should be aware that different directorates within the same large agency, such as the NSF, will have different requirements. Many librarians create data management and curation LibGuides to make this information accessible to researchers at their institution. The majority of LibGuides can be shared freely and reused by other librarians. As of this writing, over seventy data management LibGuides are available from the LibGuides Community website.[98]

Fortunately, a great deal of information is available online. The Digital Curation Centre in the United Kingdom provides a wealth of information on their website under the *Resources* and *Training* tabs, including tutorials, checklists, and briefing papers.[99] The DataOne website (Data Online Network for Earth) has a page for librarians called the Librarian Outreach Toolkit with a "Primer on Data Management," marketing materials, and talking points for discussions with researchers.[100] The *Education* tab contains ten downloadable lesson modules that cover most areas of the planning process and a database with information, and links to a variety of software applications useful at different stages of the data life cycle. DataOne also links to the Data Management Planning Tool (DMPTool), which "is a means for researchers to develop practical data management plans consistent with agency requirements and available resources."[101] In addition to being a repository for social science research data, the Inter-university Consortium for Political and Social Research (ICPSR) website has a section titled *Resources for Instructors*, including learning guides with exercises for teaching students about data, a section on data management and curation that includes suggested metadata elements for DDI metadata and information on data to exclude for confidentiality purposes, and a selection of curation tools and services.[102] Librarians should be aware that most resources provided by these organizations are targeted at researchers, but they all provide excellent education materials for librarians as well.

The Data Curation Profile website at Purdue University contains a toolkit composed of a user's guide, an interviewer's manual, an interview worksheet, and a template for creating the data curation profile.[103] The project also maintains an open-access repository of profiles for a variety of subject areas and encourages librarians to deposit and share their locally created profiles. The depth of information provided by the Data Curation Profile website is an invaluable resource for librarians just beginning to develop research data services, or a timesaver for librarians more experienced with this type of data service.

MANTRA is a free online course for research data management training offered by EDINA, the designated National Data Centre at the University of Edinburgh.[104] The self-paced curriculum consists of eight modules covering the important components of data management planning and data handling tutorials, using SPSS, R, ArcGIS, and NVivo software. Each module consists of twenty-five slides that include videos, links to resources, templates, and brief quizzes. MANTRA also includes a "Do-It-Yourself Research Data Management Training Kit for Libraries."

LC provides an online curriculum, Digital Preservation Outreach and Education. Although some of the educational and informational materials are not so much a curriculum as links to the information, the site provides a considerable amount of information about still-image and multimedia data formats, reports, and other information from the National Digital Stewardship Alliance and standards used at LC.[105]

Finally, a group of northeast research libraries, led by the University of Massachusetts Medical School Library, was funded to create the "Frameworks for a Data Management Curriculum."[106] Differing from the previously cited curricula, the framework provides an outline of the necessary components to teach data management. "The curriculum [was] designed as a series of seven course modules in order to allow maximum flexibility for customizing instruction."[107] The modules can be used together as a complete course on data management or individually to target a particular learning need. Each module consists of lesson plans, reading lists, student exercises, research data management cases, and assessments. Each lesson plan consists of learning objectives, lecture outline, activities, assessment, and readings. While many of the linked resources are health care related, the extensive curriculum is discipline neutral and can be easily adapted to any subject area.

Many more resources are available in books, in journal articles, and on the web, which cover every aspect of data management, curation, and preservation. Information on each component of the topic could fill hundreds of gigabytes of space. The broad range of disciplines makes it difficult to discuss a topic from a single point of view. Fortunately, information about data management curation and preservation is available for many specific subject disciplines. The web linking system facilitates discovery of these resources and offers another reason for the move toward contextual linked data!

CONCLUSION

The dramatic and ongoing changes in higher education and academic libraries signal paradigm shifts for the operations in technical services departments. Pressure for greater accountability by education communities filters down to the faculty and staff. The faculty is asked to bring in more research grant money, and irrespective of funding, they are expected to conduct research and involve students in the process. Providing research data services to faculty and students is an excellent and highly visible way to demonstrate the library's integration into the research life of the institution and affords an assessable metric for accountability.

The vast amount of literature on research data management indicates the growing importance of data curation and data management planning. It seems likely that the number of agencies requiring data management plans and open access to data sets will only increase. Big science research projects rely on grant funds, and researchers must work to deadlines. Several authors cited in this chapter reported that data curation was not a high priority in terms of researcher time. If graduate students are assigned this task, they will need training in data management—an area where librarians excel. Carlson reminds us that the majority of research conducted campuswide is *small* research projects operating with little funding and a small number of staff; he conjectures that this is probably true for most universities. These small research projects could be the low-hanging fruit for launching research data services.

Much of the process of data management is a collaborative effort—one in which librarians in all sectors of technical services and beyond will have future roles. Librarians in acquisitions and collection development will collaborate with researchers on acquiring existing data sets through purchase or licensing. Knowledgeable catalogers will help organize research documentation, identify standard controlled vocabularies, and provide guidance and best practices for building new taxonomies and ontologies. Metadata specialists will identify appropriate metadata standards, modify existing schemes, and create APs. Scholarly communication librarians can assist with details of depositing data sets in institutional repositories and help locate existing subject repositories. Information technologists are needed to upgrade existing technology to accommodate digital data sets and install or build applications for new data services.

Collaboration must extend outside the technical services department to other areas of the library and campus. Archivists will work with metadata specialists on digital data archiving and preservation. Reference and instruction librarians are generally the most experienced at offering information literacy classes. Instruction will extend to data literacy, which will be taught collaboratively, with each module assigned to a library expert in that particular area. Adept at the reference interview, reference and liaison librarians will be assets for interviewing researchers to develop a data curation profile. Metadata specialists, catalogers, and others involved in data management and curation will participate in planning the data curation profile interview. And harking back to Wakimoto, collaboration between the library, research computing, and the research office is critical for the success of any research data services.

Finally, everyone will have to ramp up their knowledge and skills for the specific role they will play in research data services offered by the library. It will be critical for librarians to understand the research process in specific subject areas, the flow of the data life cycle, the process of developing a data management plan, and the use of standards and application profiles. Familiarity with the Semantic Web and the linked open data structure will be essential for future catalogers and metadata librarians alike. Of course, not everyone will be expert in all areas, but collaboration among members of the research data services team will cover all areas in depth.

Research data services will not have a meteoric implementation as a full-blown program overnight. Librarians must learn what local research services are needed, and what is already available. They must decide on the types and extent of services that can be realistically offered and identify the resources needed for the program to grow. Start small and plan for the future. According to the ACRL "Environmental Scan 2013," "There will likely be a substantial role for librarians in curating, managing, and preserving data. Many predict that professional opportunities will increasingly be centered in this area through the retraining, reorganizing, and repositioning of staff." Yes, research data services will be in the future of technical services, and it will be an interesting and stimulating future indeed.

NOTES

1. Association of College and Research Libraries 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, no. 6 (2012): 312, http://crln.acrl.org.

2. Becky Yoose and Jody Perkins, "The Linked Open Data Landscape in Libraries and Beyond," *Journal of Library Metadata* 13, nos. 2–3: 197–211, doi:10.1080/19386389.2013.826075.

3. Charles W. Bailey, "Research Data Curation Bibliography," version 3, June 17, 2013, http://digital-scholarship.org/rdcb/rdcb.htm. In this 2013 version, Bailey cites 235 books and articles on data curation alone.

4. Association of College and Research Libraries Research Planning and Review Committee, "Top Ten Trends in Academic Libraries: A Review of the Trends and Issues Affecting Academic Libraries in Higher Education," *College and Research Libraries News* 75, no. 6 (2014): 294–302.

5. Ibid., 294–95.

6. Library Information and Technology Association, "Top Tech Trends—Midwinter 2014," http://www.ala.org/lita/2014midwinter. Top Tech Trends is a LITA featured session at every ALA conference.

7. Tyler Walters and Katherine Skinner, *New Roles for New Times: Digital Curation Preservation, Report Prepared for the Association of Research Libraries* (Washington, DC: Association of Research Libraries, 2011), http://www.arl.org/storage/documents/publications/nrnt_digital_curation17mar11.pdf.

8. The policy was updated to a mandate in 2008. In February 2013 the Office of Science and Technology issued a memorandum outlining a new policy on "increasing access to the results of federally funded scientific data" that included digital data sets as well as peer-reviewed publications. See http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

9. The National Information Standards Association (NISO), *Recommended Practices for Online Supplemental Journal Article Materials*, January 2013, http://www.niso.org/apps/group_public/download.php/10055/RP-15-2013_Supplemental_Materials.pdf. The recommendation covers accompanying data sets, metadata, and metadata for objects stored in local repositories.

10. White House, "Executive Order: Making Open and Machine Readable the New Default for Government Information," press release, May 9, 2013, http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government.

11. *Open Data Charter*, https://www.gov.uk/government/publications/open-data-charter. The five Principles are: (1) open data by default; (2) quality and quantity; (3) usable by all; (4) releasing data for improved governance; and (5) releasing data for innovation.

12. United States Government, *U.S. Open Data Action Plan*, May 9, 2014, http://www.whitehouse.gov/sites/default/files/microsites/ostp/us_open_data_action_plan.pdf.

13. Clifford A. Lynch and Joan K. Lippincott, "Institutional Repository Deployment in the United States as of Early 2005," *D-Lib Magazine* 11, no. 9 (September 2005): doi:10.1045/september2005-lynch.

14. Anna Gold, "Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians," *D-Lib Magazine* 13, nos. 9–10 (2007), doi:10.1045/september2007-gold-pt1; "Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries," *D-Lib Magazine* 13, nos. 9–10 (2007), doi:10.1045/september2007-gold-pt2.

15. Graham Pryor, ed., *Managing Research Data* (London: Facet Publishing, 2012), 2–3.

16. E. Deumens et al., "Research Data Lifecycle Management: Tools and Guidelines," June 17, 2011, http://ufdc.ufl.edu/IR00000570/00001.

17. Erin Bartolo and Jill Hurst-Wahl, "Curating Library Data, Big Data, Data Sets: Using Big Data for Library Advocacy," Syracuse University School of Information Studies, September 27, 2013, http://www.slideshare.net/jill_hw/curating-library-data-big-data-data-sets?qid=d0642ffe-dde6-447a-9b46-2c64646cc852&v=default&b=&from_search=1.

18. 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, no. 4 (2011): 915–37, doi:10.1353/pla.2011.0049.

19. Ibid.

20. Ibid., 922.
21. Gail Steinhart et al., "Prepared to Plan? A Snapshot of Researcher Readiness to Address Data Management Planning Requirements," *Journal of eScience Librarianship* 1, no. 2 (2012): 63–78, doi:10.7191/jeslib.2012.1008.
22. Ibid., 77.
23. Lori Jahnke, Andrew Asher, and Spencer D. C. Keralis, *The Problem of Data* (Washington, DC: Council on Library and Information Resources, 2012), 15.
24. Carol Tenopir, B. Birch, and S. Allard, "Academic Libraries and Research Data Services: Current Practices and Plans for the Future," white paper, Association of College & Research Libraries, June 2012, http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf.
25. Jina Choi Wakimoto, "Developing Research Data Management Services," *EDUCAUSE Review Online*, February 26, 2013, http://www.educause.edu/ero/article/developing-research-data-management-services.
26. Ibid.
27. Tenopir, Birch and Allard, "Academic Libraries and Research Data Services," 8.
28. ACRL Research Planning and Review Committee, "Environmental Scan 2013," April 2013, http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/EnvironmentalScan13.pdf.
29. Of the 330 articles published between January 2005 and June 2014 on data curation that are cited in Bailey's "Bibliography," 73 percent were published between January 2011 and June 2014, a period of only three and a half years.
30. David Fearon Jr., Betsy Gunia, and Barbara E. Pralle, *SPEC Kit 334: Research Data Management Services* (Chicago: Association of Research Libraries, 2013).
31. University of California Curation Center, Data Management Planning Tool, https://dmp.cdlib.org/index.html.
32. Fearon, Gunia, and Pralle, *Spec Kit 334*, 12.
33. Tenopir, Birch, and Allard, "Academic Libraries and Research Data Services," 3.
34. Mark Dahl, "Data-Driven Liberal Arts: The Library Role," *Academic Commons: For the Liberal Education Community*, July 24, 2014, http://www.academiccommons.org/2014/07/24/data-driven-liberal-arts-the-library-role/; Megan Toups and Michael Hughes, "When Data Curation Isn't: A Redefinition for Liberal Arts Universities," *Journal of Library Administration* 53, no. 4, (2013): 223–33, doi:10.1080/01930826.2013.865386; Tenopir, Birch, and Allard, "Academic Libraries and Research Data Services," 3.
35. Katherine Goold Akers, "Looking Out for the Little Guy: Small Data Curation," *Bulletin of the American Society for Information Science and Technology* 39, no. 3 (2013): 58–59.
36. Jake R. Carlson, "Demystifying the Data Interview: Developing a Foundation for Reference Librarians to Talk with Researchers about Their Data," *Reference Services Review* 40, no. 1 (2012): 7–23.
37. Akers, "Looking Out for the Little Guy."
38. Toups and Hughes, "When Data Curation Isn't."
39. Jahnke, Asher, and Keralis, *The Problem of Data*, 8; Jessica Adamick, Rebecca C. Reznik-Zellen, and Matt Sheridan, "Data Management Training for Graduate Students at a Large Research University," *Journal of eScience Librarianship* 1, no. 3 (2012), dx.doi.org/10.7191/jeslib.2012.1022.
40. Angela Brew, "Understanding the Scope of Undergraduate Research: A Framework for Curricular and Pedagogical Decision-Making," *Higher Education* 66, no. 5 (2013): 603–18, dx.doi.org/10.1007/s10734-013-9624-x; Pete Smith and Chris Rust, "The Potential of Research-Based Learning for the Creation of Truly Inclusive Academic Communities of Practice," *Innovations in Education and Teaching International* 48, no. 2, (2011): 115–25, dx.doi.org/10.1080/14703297.2011.564005.
41. Council on Undergraduate Research (CUR), "Conferences and Events," http://www.cur.org/conferences_and_events/.
42. Dahl, "Data-Driven Liberal Arts."
43. Data Curation Centre, "DCC Charter and Statement of Principles," http://www.dcc.ac.uk/about-us/dcc-charter/dcc-charter-and-statement-principles.

44. Sarah L. Shreeves and Melissa H. Cragin, "Introduction: Institutional Repositories: Current State and Future," *Library Trends* 57, no. 2 (2008): 93, dx.doi.org/10.1353/lib.0.0037.

45. Gold, "Cyberinfrastructure, Data, and Libraries, Part 1."

46. Anna Gold, "Data Curation and Libraries: Short-Term Developments, Long-Term Prospects," Robert E. Kennedy Library, April 4, 2010, http://digitalcommons.calpoly.edu/lib_dean/27/.

47. Committee on Earth Observational Satellites Working Group on Information Systems and Services (WGISS), *Data Life Cycle Models and Concepts*, CEOS 1.2, Version 13.0, April 19, 2012, http://ceos.org/ourwork/workinggroups/wgiss/interest-groups/data-stewardship/.

48. John L. Faundeen et al., "The United States Geological Survey Science Data Lifecycle Model," U.S. Geological Survey, Open-File Report 2013–1265, 2013, dx.doi.org/10.3133/ofr20131265.

49. University of Virginia, Research Data Services, "Research Data Life Cycle," 2014, http://dmconsult.library.virginia.edu/.

50. Andrew Treloar, David Groenewegen, and Catherine Harboe-Ree, "The Data Curation Continuum: Managing Data Objects in Institutional Repositories," *D-Lib Magazine* 13, nos. 9–10 (2007), http://www.dlib.org/dlib/september07/treloar/09treloar.html.

51. Merinda McClure et al., "Data Curation: A Study of Researcher Practices and Needs," *portal: Libraries and the Academy* 14, no. 2 (2014): 151–52, dx.doi.org/10.1353/pla.2014.0009.

52. Melissa H. Cragin et al., "Data Sharing, Small Science and Institutional Repositories," *Philosophical Transactions of the Royal Society A* 368 (2010): 4023–38, doi:10.1098/rsta.2010.0165.

53. Ibid.

54. McClure et al., "Data Curation."

55. See the DataCite website at http://www.datacite.org/.

56. Force11, Data Citation Synthesis Working Group, "Joint Declaration of Data Citation Principles," https://www.force11.org/datacitation.

57. "DataCite Metadata Schema for the Publication and Citation of Research Data," version 3.0, July 2013, doi:10.5438/0008.

58. Jake Carlson, "Data Management Plan Self-Assessment Questionnaire," Purdue University Libraries, 2011, https://purr.purdue.edu/dmp/self-assessment.

59. Michael Witt et al., "Constructing Data Curation Profiles," *International Journal of Digital Curation* 4, no. 3 (December 2009): 93–103, dx.doi.org/10.2218/ijdc.v4i3.117.

60. Ibid., 93.

61. Ibid., 95.

62. Jake Carlson and D. Scott Brandt, *Data Curation Profiles Directory*, http://docs.lib.purdue.edu/dcp/.

63. Sherry L. Vellucci, "Knowledge Organization," in *Academic Library Research: Perspectives and Current Trends*, ed. Marie L. Radford and Pamela Snelson, 138–88 (Chicago: Association of College and Research Libraries, 2008).

64. Carlson, "Demystifying the Data Interview."

65. Jane Greenberg, "Metadata for Managing Scientific Research Data," presentation, NISO/DCMI Webinar, August 22, 2012, http://www.niso.org/news/events/2012/dcmi/scientific_data/.

66. Each of these types of metadata are described in detail and discussed in Steven J. Miller, *Metadata for Digital Collections: A How-to-Do-It Manual* (New York: Neal Schuman, 2011).

67. United States Environmental Protection Agency, "Data Standards: Fact Sheets," http://iaspub.epa.gov/sor_internet/registry/datastds/outreachandeducation/educationalresources/factsheets/.

68. International Standards Organization, "ISO 19115-1:2014, Geographic Information—Metadata—Part 1: Fundamentals," http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798.

69. See the Data Documentation Initiative website at http://www.ddialliance.org/.

70. Digital Curation Centre, "Disciplinary Metadata," http://www.dcc.ac.uk/resources/metadata-standards.

71. JISC, "Putting Things in Order: A Directory of Metadata Schemas and Related Standards," http://www.jiscdigitalmedia.ac.uk/guide/putting-things-in-order-links-to-metadata-schemas-and-related-standards.

72. Society of American Archivists, "Metadata Directory," http://www2.archivists.org/groups/metadata-and-digital-object-roundtable/metadata-directory.

73. Jenn Riley, "Seeing Standards: A Visualization of the Metadata Universe," http://www.dlib.indiana.edu/~jenlrile/metadatamap/.

74. NASA, "Metadata Protocol and Standards," http://gcmd.nasa.gov/add/standards/index.html.

75. NOAA, National Geophysical Data Center, "U.S. ECS Project Data Management," http://www.ngdc.noaa.gov/mgg/ecs/metadata/.

76. EPA, "Metadata Standards," http://systemofregistries.supportportal.com/link/portal/23002/23017/ArticleFolder/1581/Metadata-Standards.

77. Springshare, LibGuides Community, http://libguides.com/community.php.

78. Miller, *Metadata for Digital Collections*, 260.

79. ANSI/NISO Z39.19-2005 (R2010), *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, http://www.niso.org/apps/group_public/download.php/12059/z39-19-2005r2010.pdf.

80. Ibid.

81. NASA, *NASA Thesaurus, Volume 1: Hierarchical Listing with Definitions*, January 2012, http://www.sti.nasa.gov/thesvol1.pdf; and *Volume 2: Rotated Term Display*, January 2012, http://www.sti.nasa.gov/thesvol2.pdf.

82. Getty Research Institute, *Art & Architecture Thesaurus Online*, http://www.getty.edu/research/tools/vocabularies/aat/.

83. Lars Marius Garshol, "Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of It All," *Ontopia*, October 26, 2004.

84. Ibid.

85. Klaar Vanopstal et al., "Vocabularies and Retrieval Tools in Biomedicine: Disentangling the Terminological Knot," *Journal of Medical Systems* 35, no. 5 (2011): 527–43, dx.doi.org/10.1007/s10916-009-9389-z.

86. Ibid.

87. Thomas R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," *International Journal of Human-Computer Studies* 43, nos. 5–6 (1995): 907–28.

88. Vanopstal et al., "Vocabularies and Retrieval Tools," 533.

89. Steven J. Miller, "Introduction to Ontology Concepts and Terminology," DC-2013 Tutorial, September 2, 2013, http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/140/105.

90. Dublin Core Metadata Initiative, "DCMI Abstract Model," http://dublincore.org/documents/2007/06/04/abstract-model/.

91. Dublin Core Metadata Initiative, "Guidelines for Dublin Core Application Profiles," http://dublincore.org/documents/profile-guidelines/#sect-1.

92. Miller, *Metadata for Digital Collections*.

93. Ibid.

94. Library of Congress, "LC Linked Data Services: Authorities and Vocabularies," http://id.loc.gov/.

95. Silvia B. Southwick and Cory K. Lampert, "Not Just for Geeks: A Practical Approach to Linked Data for Digital Collections Managers" Mountain West Digital Library, October 2013, http://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=1116&context=libfacpresentation.

96. OCLC, "Linked Data Survey Results—1: Who's Doing It (Updated)," August 28, 2014, Hangingtogether.org, http://hangingtogether.org/?p=4137.

97. Pryor, *Managing Research Data*, 10.

98. See the LibGuides Community website at http://libguides.com/community.php?m=g.

99. Digital Curation Centre, "Roles," http://www.dcc.ac.uk/resources/roles.

100. DataOne, "Librarians Outreach Toolkit," https://www.dataone.org/for-librarians.

101. DataOne, "Investigator Toolkit," https://www.dataone.org/investigator-toolkit.

102. See the Inter-university Consortium for Political and Social Research website at http://www.icpsr.umich.edu/icpsrweb/landing.jsp.

103. Data Curation Profiles Toolkit, http://datacurationprofiles.org/.

104. MANTRA, Research Data Management Training, http://datalib.edina.ac.uk/mantra/.

105. Library of Congress, "Digital Preservation Outreach and Education," http://www.digitalpreservation.gov/education/curriculum.html.

106. "Frameworks for a Data Management Curriculum," http://library.umassmed.edu/data_management_frameworks.pdf.

107. Ibid., 2.