CHAPTER 6

Data Sharing in the Sciences

Stacy Kowalczyk and Kalpana Shankar
Indiana University, Bloomington, IN

Introduction

Collaboration across disciplines, the increasing size of those collaborations, and the application of information technologies to scientific problems have contributed greatly to the data-intensive nature of scientific research. These trends have reduced the relevance of distance, allowed scientists to share resources remotely, and increased both the complexity and quantity of research data. They have also created numerous challenges for the sharing and reuse of data through data repositories. When institutions create repositories, they do so with several goals in mind: grant fulfillment, peer review, long-term archiving, disease management, daily scientific practice, the leveraging of scarce or endangered resources, and the exploiting of large infrastructures.

Tools for data mining and analysis foster the integration of existing datasets to answer and promote new lines of inquiry, enable multidisciplinary research, and facilitate educational efforts. Data may be archived within the research group, by an external organization, or in an institutional repository. Increasingly, it is not just researchers who reuse data, but also educators, policymakers, and the general public. Making data broadly available can promote public understanding of science, evidence-based advocacy, educational uses, and citizen-science initiatives. Promoting the effective sharing of data is increasingly part of national and international scientific discourse and held to be essential to the future of science (Interagency Working Group on Digital Data, 2009). Funding agencies argue that data sharing promotes the wisest use of public resources by reducing repetitive collection of expensive or sensitive data. Examples include datasets collected in fragile areas of the world, with national security implications, from unique circumstances such as natural or human-made disasters, or composed of rare or complex samples. In short, data sharing furthers the spirit of inquiry with a focus on interoperability and reuse by making experimental and observational data available (at least to the scientific community) and thus verifiable and replicable.

National initiatives to develop both an infrastructure to facilitate data sharing and a culture of sharing are gradually emerging (Research

Information Network, 2008) but the sharing of datasets is not simple nor is it practiced across all scientific disciplines. Data sharing is not universally practiced for numerous reasons. Organizations, as well as individual researchers, may consider themselves to be either friendly competitors or rivals for funding, leading them to view their data as a source of competitive advantage (Chaplin, 2004). Data and other by-products of research represent important and sometimes invaluable intellectual capital; dissemination strategies should be synchronized with the mission of the organization that collected the data (Association of Research Libraries, Association of American Universities, Coalition for Networked Information, & National Association of State Universities and Land-Grant Colleges, 2009). Individuals and organizations may fear being perceived as incompetent or inefficient if they have not extracted the full economic and institutional value from the data they share; further, organizations worry that trustees, donors, and other shareholders would not approve of unrestricted access that allows others to "free-ride" on institutional data resources (Hammond, Moritz, & Agosti, 2008, p. 3). Another concern is that data sharing may open up researchers to criticism of data collection methods and their research in general.

Minimally, for data sharing to be effective, the datasets must satisfy three criteria: persistence, longevity and sustainability, and quality (Interagency Working Group on Digital Data, 2009; Klump, Bertelmann, Brase, Diepenbroek, Grobe, Höck, et al., 2006). In other words, the data (and their location) must be findable over time via consistent pointers, preserved and accessible for the long term, and of sufficient quality to be usable. To meet these criteria, data repositories require the development of appropriate technical and organizational infrastructures for storage and retrieval, incentives for researchers to deposit and use the data, and enforceable policies. (Research groups and projects, academic and research institutions, funding agencies, and disciplinary bodies and societies may all have relevant data sharing policies.) Although the technologies exist to archive large datasets, making them readily accessible and easily described is not a trivial undertaking.

Many disciplines have successfully encouraged data sharing through its integration into the publication and peer review process (Piwowar & Chapman, 2008). For example, the Inter-university Consortium for Political and Social Research (ICPSR) has been active in archiving social science datasets, migrating them as necessary, and providing training for researchers from around the world in their use. In some disciplines, archiving data is a prerequisite for publication. A recent survey of the policies of journals in a subdiscipline of biology found that three quarters of the journals surveyed had at least some mention of a data sharing requirement for publication in the journal (Piwowar & Chapman, 2008). However, such policies do not always seem to work uniformly or effectively; one survey found that 28 percent of genetics researchers were unable to confirm published results because they could not access the relevant data (Campbell, Clarridge, Gokhale, Birenbaum,

Hilgartner, Holtzman, et al., 2002); many junior scientists reported in a 2006 survey that they had experienced data withholding (Vogeli, Yucel, Bendavid, Jones, Anderson, Louis, et al., 2006). In short, responsibility for making data available varies widely from discipline to discipline, and within disciplines, with differing formats, practices, and costs (and bearers of those costs).

For the information professions, several practical, significant questions arise around data sharing. These are intersections of storage, retrieval, preservation, management, access, and policy issues. The design and deployment of storage and retrieval mechanisms, metadata schema for data description, long-term preservation for data archives, the education of data curators and users, development of institutional and organizational policy, and access and use of repositories are some of the problems with which information professionals engage. More broadly, data sharing raises questions about the nature of research itself—the role of the public and the private sectors, citizen participation in the heavily taxpayer-supported scientific process, and the equitable distribution of the results of the research enterprise. Taken together, these disparate, complex challenges constitute the *data sharing problem*.

In this chapter we explore the many strands of data sharing and its myriad technical, social, ethical, and policy challenges. We focus specifically on the natural sciences, although many of the issues we discuss may be of equal relevance to the social sciences and humanities. We discuss the nature of data and the life cycle of creation and use, scale, formats, and provenance and the research problems these issues present. We then explore the policy, technical, ethical, and organizational dimensions of data sharing, present several case studies of data sharing in action, and conclude with an overview of current and future directions, including the use of Web 2.0 for data sharing and archiving.

Structure of the Chapter

Data sharing crosses many disciplines, professions, and literatures; at its core it is a sociotechnical enterprise. In framing this chapter, we found that this divergence presents a number of structural problems for both the writer and the reader in exploring the issues. Thus, we introduce a framework with which we explore the issues, noting that we could have presented the same information in many other ways.

We begin by discussing the nature of data—formats, the context in which they are created, measures of their quality, and their persistence (or lack thereof). We then explore aggregations of data into repositories, or data collections, and the nature of those aggregations.

Not surprisingly, the technical needs of data sharing command a considerable amount of resources, human and otherwise. We call this layer the technical infrastructure. The technical infrastructure for data sharing is a rich field of research. Most of the communities that have built data sharing infrastructures have developed unique solutions

(Rajasekar, Marciano, & Moore, 1999). Although this may seem wasteful, and some approaches undoubtedly are, there is no single infrastructure that will serve all of the varying needs of the scientific community. Infrastructure, which encompasses hardware, software, data formats, and protocols, needs to be built, implemented, and evaluated based on how the data are to be used (Geschwind, 2001). For technical infrastructure, we explore the three minimal criteria that Klump and colleagues (2006) articulate for data sharing: persistence, longevity, and quality. Data should be described, stored, accessible, and secure for these criteria to be applied.

The major interactions that need to be addressed in a technical infrastructure for data sharing include data discovery, data access, data context, data integrity, data security, and data management. We refer to this as the access and discovery layer. Descriptive standards, metadata, vocabularies, the context of creation and storage, and identifiers for data and their locations are critical to making data persistent, discoverable, and accessible.

The institutional, professional, and social conditions under which data are created and shared permeate these other layers, but can be most clearly explicated separately from the technical dimensions of data sharing. Many researchers are understandably concerned about the overhead entailed in making data publicly available, having their data used by other researchers before the data producers publish their findings (Toronto International Data Release Workshop, 2009), and the ways that outside (i.e., outside of the researcher's workgroup) entities manage data. They may also be concerned about the security of sensitive data in institutional repositories. Research ethics, disciplinary norms, complex institutional networks (including private and public partnerships), and resources (human and otherwise) for making data available can serve as either barriers or incentives for data sharing. Access policies and mechanisms also need to be developed that appropriately reflect the value of research data, particularly in the university setting. At this human infrastructure layer, explicit concerns about intellectual property, ethics and privacy, access and use policies, business models, sustainability, scientific practice, and professionalism arise. We conclude with a discussion of evaluation and suggest future research directions.

The Nature of Data

Defining Data

Data sharing begins with, of course, data. Although research data can be generated digitally, on paper, or in some hybrid form, most of the literature on data sharing focuses on data that is generated digitally. We generally adopt the definition of data as put forth by the U.S. National Science Board (NSB) (2005, p. 9): "any information that can be stored in digital form, including text, numbers, images, video or movies, audio,

software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment." Hilgartner (1995) argues that data should be construed more broadly because conflicts over access have included tangible artifacts such as biological reagents and molecules, instrumentation, and other key resources. Including these artifacts is beyond the scope of this chapter. Data are inherently collective and come in sets—the collation of many individual data. In most scientific research, a researcher would need a number of datasets or databases. We refer to these sets of datasets as data collections.

Data collections "refer not only to stored data but also to the infrastructure, organizations, and individuals necessary to preserve access to the data" (U.S. National Science Board, 2005, p. 9). The NSB has developed a hierarchy by size of collection types: research data collections, community data collections, and reference data collections. We acknowledge that this hierarchy is not used by data creators themselves but, because this language is prevalent, we use the terms to structure discussion.

The process by which data are created, analyzed, and managed is complex; the experimental and observational data collection process is often referred to as the data life cycle. In general, although there may be significant differences in the individual stages, the life cycle is assumed to encompass the experimental design and capture, cleaning/integration, analysis, publication, and preservation processes, which occur in an iterative fashion.

The life cycle approach has been recognized as important to data preservation and sharing because active intervention early in that life cycle is essential for successful preservation (Beagrie, 2006). A high-level hypothesis or research question should be broken down into tasks that can be performed with appropriate workflow and data analysis. In e-science, such experiments are conducted *in silico* (performed computationally) but the same life cycle exists (Wroe, Goble, Greenwood, Lord, Miles, Papay, et al., 2004). However, different kinds of data (and disciplines) have different life cycles. Wallis, Borgman, Mayernik, and Pepe (2008) argue that the number of individuals and institutions involved at each stage of the life cycle increases as the complexity of the data increase. The cumulative weight of decisions made at each stage determines what is available at the next stage, how it is handled, and the purposes for which it is useful. Although the data themselves may be more or less easily depicted through various descriptive processes, documenting the decisions at each stage of the life cycle, and describing that cycle in detail, is more problematic and not easily automated (Borgman, 2007; Higgins, 2008; Wallis et al., 2008).

Format

Scientific data can be numeric, text-based, audio, video, or still images, depending on the nature of the studies that generated them. Data may

be represented in a variety of formats that are domain- and community-dependent. Many communities have developed a set of standard formats to facilitate data sharing and system interoperability (Gardner, Toga, Ascoli, Beatty, Brinkley, Dale, et al., 2003; Westbrook, Ito, Nakamura, Henrick, & Berman, 2005). Researchers in astrophysics and genomics, for example, have a known set of widely accepted file formats that facilitate both data sharing and data archiving (Interagency Working Group on Digital Data, 2009).

Since the 1990s, the scientific community has been aware that effective data sharing depends on standard data formats and tool sets to build, read, and manage the data (Gardner et al., 2003; Rew & Davis, 1990). During the research life cycle, the original, raw data are often processed and transformed into a reduced or derived data product (Research Information Network, 2008). The wide variety of data formats used in different scientific disciplines is a barrier to data mining and sharing; conversion between different formats requires a substantial effort (Mann, Williams, Atkinson, Brodlie, Storkey, & Williams, 2002; Shiffrin & Börner, 2004). The ease or difficulty of data reuse, as well as the probability of long-term access and persistence, depends on the complexity and transparency of the file format itself (Abrams, 2004). Because formats depend on specific technologies and are bound to both a software environment and a hardware infrastructure for rendering and processing, changes to the format or to the environment can make the underlying data unrecognizable and unusable. One study, not surprisingly, discovered that it was easier to convert data in open formats such as Tagged Image File Format (TIFF). But the researchers also discovered that even open formats can have a proprietary, and thus secret, set of tags that may not be supported in future versions; they may also lack documentation for migration or conversion (Lawrence, Kehoe, Rieger, Walters, & Kenney, 2000). Because they manage and archive large collections of heterogeneous digital files, the Library of Congress; the National Archives of England, Wales and the United Kingdom; OCLC; and the National Archives of Australia have each developed separate sets of criteria for assessing the risks associated with formats. Although these four criteria differ in terms of complexity and thoroughness, a consensus exists that archival formats should be well documented and well understood; not wholly owned by a single commercial entity; widely adopted to increase the probability of commercial tools for migration and to ensure a long usage cycle to avoid repeated, short-term migration; self-contained; and not reliant on a specific technical environment (Kowalczyk, 2007).

Context

Context is defined as the information that documents the relationships of the data content to its environment (Consultative Committee for Space Data Systems [CCSDS], 2002). Context can include descriptions

of the purpose of the data, its quality, formats, and many other variables. In short, context documents how datasets fit into their physical and technical environments (file formats and field descriptors) as well as into the scientific environment (experiment treatments and applications). Representing context—creating the relevant metadata—is as essential as representing datasets themselves. Moore, Baru, Rajasekar, Ludaescher, Marciano, Wan, and colleagues (2000a) contend that context, maintained in a representation that is not dependent upon the format of the data it describes, is a significant component of a persistent data infrastructure. As part of a project to develop a data sharing environment for a community of biologists, Chin and Lansing (2004) describe the context necessary for successful collaboration: biologists need to understand the relationships between the data and software applications, experiments, projects, and the scientific community. The specific properties of the data that should be included in the context are: general dataset properties (such as owner, creation date, size, and format); experimental properties (conditions and properties of the scientific experiment that generated or is to be applied to the data); data provenance (the relationship of data to previous versions and other data sources); integration (the relationship of data subsets within a full dataset); analysis and interpretation (such as notes, experiences, interpretations, and knowledge generated from analysis of data); physical organization (the mapping of datasets to physical storage structures such as a file system, database, or some other data repository); project organization (the mapping of datasets to project hierarchy or organization); scientific organization (the mapping of datasets to some scientific classification, hierarchy, or organization); task (the research task(s) that generated or applies to the dataset); experimental process (the relationship of data and tasks to the overall experimental processes); and user communities (different organizations' applications of the datasets) (Chin & Lansing, 2004).

These properties are represented in specific standardized formats, often unique to each scientific domain. Some are semi-structured representations for the digital objects and the data collection using an open schema definition (Moore, Baru, Rajasekar, Ludaescher, Marciano, Wan, et al., 2000a, 2000b). Some are multilevel representations with dataset level and variable/attribute level metadata (Jaiswal, Giles, Mitra, & Wang, 2006). Others are represented as graphs (Chen & Kotz, 2001), as topic maps (Goslar & Schill, 2004), as semantic networks (Henricksen, Indulska, & Rakotonirainy, 2002), or as trees (Gardner, Goldberg, Grafstein, Robert, & Gardner, 2008).

Creating and documenting this context is usually a manual effort. The researcher who creates the data can also create the metadata needed for preservation in a repository. Traditionally, this process has been a one-time event. However, new social networking paradigms of user-contributed metadata are emerging; collections are beginning to ask subsequent data users to add information via textual annotations

(Chin & Lansing, 2004; Ives, Halevy, Mork, & Tatarinov, 2003; Jaiswal et al., 2006; Michener, Beach, Bowers, Downey, Jones, Ludaescher, et al., 2005; Myers, Allison, Bittner, Didier, Frenklach, Green, et al., 2005) as well as visual annotations over images (Chin & Lansing, 2004). Research communities are developing algorithms and processes to generate dynamic community vocabularies and ontologies automatically, from both the human-generated metadata and the data themselves, for data integration and discovery in data collections (Jaiswal et al., 2006). The goals of automated context creation are to generate more accurate and consistent data, to create sufficient context for precise data discovery, and to ease the burden of creating metadata from the contributing scientist (Michener, 2005).

Data Integrity

Data integrity is the expectation of data quality—that is, the assurance that the data are whole or complete, consistent, and correct. Data integrity includes both intellectual and technical wholeness and can be compromised by human errors at any stage of creation and use. Because correcting data is always expensive, the best practice for all information systems is to ensure that the data are correct from the beginning (Galloway, 2004). Ensuring the integrity of the data is a significant factor in establishing trust within the designated community and the level of use of a data collection (Lyon, 2007).

Green and Gutmann (2007) advocate for collaboration among local, institution-based repositories, the data collection repositories, and the researcher that would both simplify and enhance the quality and the integrity of the data. Their model of the research data life cycle (discovery and planning, initial data collection, final data preparation and analysis, and publication and sharing) is designed to optimize persistent access to data. By involving the repository in all phases of the research project, from conceptualization and grant writing through publication, they argue that integrity issues such as data creation rules, processing requirements, data standards, and data confidentiality can be addressed early, creating a set of data and metadata with integrity and quality that is ready for long term archiving within repositories (Green & Gutmann, 2007).

Fixity is a significant data integrity issue because digital data can be changed easily, either maliciously or inadvertently (Hedstrom & Montgomery, 1998). Gladney (2004) argues that to ensure integrity the repository must be able to guarantee the fixity of each object. Technically, this is a simple process that includes creating a checksum or a digital signature for each object, be it a data file, a configuration file, rendering software, or documentation. The checksum should be created before the data are placed into the repository, stored in the repository with administrative data about the data object, and validated by a process that calculates the checksum for each digital file and compares

it to the saved checksum on a regular schedule, reporting errors to the repository managers (Kowalczyk, 2007).

Versioning

Discussions of data sharing have generally focused on final data. The National Institutes of Health (NIH) defines "final data" as "recorded factual material commonly accepted in the scientific community as necessary to validate research findings" (National Institutes of Health, 2003, para. 4). The NIH accepts only data that support publications and explicitly excludes paper laboratory notebooks and other artifacts that document work in progress, even though these represent much of the data generated in the conduct of science. Much of the emphasis on data sharing thus focuses on finished or finalized datasets that can be used by groups and individuals beyond the original creators. However, it is not always clear which data should be considered final. Treloar, Groenewegen, and Harboe-Ree (2007) describe a fluid state for data, a continuum between informal pre-publication and more formal publication. Steinhart (2007) asserts that researchers need to share intermediate and supporting data and proposes that libraries provide both staging repositories, to allow researchers to add, delete, modify and share as they see fit prior to publication, and a process for moving completed datasets to an archival repository.

In some fields it is not sufficient to share the raw datasets; descriptors and the tools and methods by which data are generated, analyzed, and shared must be made available if the data are to be fully useful. The National Aeronautics and Space Administration (NASA) (1986), in its work developing earth observation data systems, developed a hierarchy of data levels to describe the nature of the data: Level 0, unprocessed instrument data at full resolutions; Level 1A, unprocessed instrument data at full resolution, time referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and geo-referencing parameters; Level 1B, which are Level 1A data that have been processed within the range of the sensor filters; Level 2, derived environmental variables at the same resolution and location as the Level 1 source data; Level 3, variables mapped to uniform special and temporal grid scales, usually with some completeness and consistency properties; Level 4, model output or results from analyses of lower-level data (i.e., variables that were not measured by the instruments but instead are derived from these measurements). These levels are being reinterpreted in new domains as data provenance descriptors (Bose & Frew, 2005) and as text encoding levels (Renear, Dolan, Trainor, & Cragin, 2009).

Data Persistence

Various terms have been used to describe the process needed to maintain digital materials over time, including digital preservation, digital

archiving, and digital curation. These terms are still evolving, are often used interchangeably, and are defined as "actions needed to maintain digital research data and other digital materials over their entire life cycle and over time for the current and future generations of users" (Beagrie, 2006, p. 4). Because persistence is a significant aspect of data sharing (Interagency Working Group on Digital Data, 2009; Klump et al., 2006), data archiving and curation are vital services of data collections (Green & Gutmann, 2007). Persistence through data curation requires organizational commitment and ongoing stewardship (Lynch, 2003). As an organizational issue, persistence requires long-term commitment for staffing, storage, and technology upgrades; thus persistence requires sustainable funding (Association of Research Libraries et al., 2009). Digital preservation policy development is a significant indicator of an organization's support for preservation activities (McGovern, 2007).

Moore (2008) characterizes digital preservation as a method of communication with both the future and the past. The conversation with the future conveys preservation properties such as authenticity, integrity, and maintenance of an object's chain of custody, so that future systems, as yet unimagined, will be able to interpret and display the information. The conversation with the past provides the characterizations of prior preservation processes and preservation management policies.

Assessing and managing risk are significant components of preservation management policy development. Kenney, McGovern, Botticelli, Entlich, Lagoze, Payette, and colleagues (2002) have identified four risk management stages: data gathering (risk identification) and characterization; simple risk declaration and detection; contextualized risk declaration and detection; and automated preservation policy enforcement. Digital preservation risk assessment methodologies include Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) and INvestigation of FOrmats based on Risk Management (INFORM). The DRAMBORA methodology has three stages: identification of preservation objectives, identification of risks along with the probabilities and impact of the risks, and identification of risk mitigation strategies and measures (Whyte, Job, Giles, & Lawrie, 2008). The INFORM methodology defines six classes of risk: digital formats, software, hardware, associated organizations, the digital repository, and format migration preservation plans (Stanescu, 2005).

Data Collections

The U.S. National Science Board (2005) broadly defines data collections as the infrastructure, organizations, and individuals needed to provide persistent access to stored data. They developed a three-layer typology of data collections organized by size and scope: research data collections, community data collections, and reference data collections. Research data collections refer to the output of a single researcher or lab during

the course of a specific research project. This collection may or may not use the data standards of its community or have use beyond its own original purpose. Community data collections generally serve a well defined arena of research. Often, standards are developed by the community to support the collection. At the highest level, reference data collections are broadly scoped, widely disseminated, well funded collections that support the research needs of many communities. In general, much of the focus of data research and development (and funding organizations' interest) has been on the trajectory of moving from local, somewhat informal research collections to more established, sustainable ones that can serve broader groups over time.

In the next section we use this framework—research collections, community collections, and reference collections—to describe the functions, structure, and organizational dimensions of such of datasets. We acknowledge that many contributors to and users of data collections generally do not use these terms to describe their own activities and repositories. Nevertheless, the framework provides a useful way of describing and categorizing many of the aspects of data collections and the changes that arise from scaling up and expanding the scope of use.

Research Collections

Research collections, the output from an individual researcher or project, are most often shared using technologies that are not specifically designed for sharing scientific data. These technologies include personal websites, email, electronic mailing lists, and online discussion groups (Gardner et al., 2003). Discovery of research collections is usually ad hoc, based on personal relationships or personal knowledge of reputation or published works (Chin & Lansing, 2004). The quality of the metadata for research collections varies widely from standardized schema to idiosyncratic labeling (Research Information Network, 2008). Data access is via simple technologies such as email attachments, File Transfer Protocol (FTP) (either secured or unsecured) or HyperText Transfer Protocol (HTTP) download. In this informal sharing environment, data security can be easily negotiated between the parties; however, determining and enforcing compliance with those security agreements can be difficult (Gardner et al., 2003). In research collections, the understanding of both data context and data integrity can benefit from the ad hoc, personal nature of the data exchange. The context of the data, as has been discussed, is a complicated interchange of implicit and explicit metadata. Direct communication between the data creator and the data user can facilitate efficient knowledge transfer. Personal relationships with the data creator and scholarly reputations typically increase confidence in the integrity of the data. Good data management practices vary widely for research collections due to a set of complex interactions among the technology environment of the academic institution and the nature of the data context, content, and format.

A new trend is emerging that might provide a more scalable model for data sharing for research collections. Academic institutions are beginning to develop programs to include data within their institutional repositories (Lyon, 2007; Parr & Cummings, 2005; Ruusalepp, 2008). Currently, adoption of institutional repositories as an infrastructure for storing and disseminating scientific data is not widespread (Lyon, 2007) but, as the barriers to self-archiving are reduced by streamlining workflows and simplifying metadata capture, researchers may find a more formal model useful to share their collections of data (Crow, 2002; Lyon, 2007).

Community Collections

Community collections integrate local, research level collections into more structured forms with the resources needed to undertake the task. They often face multiple challenges of standardization of data formats, metadata, and harmonizing diverse data management practices. To take one example of such a transitional community collection, coastal resource assessment and management relies heavily on numerous computer applications and data sources that are created and used by local, state, and federal government agencies; academic units; and nonprofit organizations. To manage coastal resource information and integrate data from multiple organizations, a partnership was formed among entities in the state government, Oregon State University, and a regional nonprofit to produce a data portal entitled the Oregon Coastal Atlas. This portal was designed to provide access to datasets, mapping tools, and other resources for decision making and research by interested parties. However, even experienced creators and users of the Atlas's data have had serious problems finding and using relevant data. They are confounded by multiple versions of the same data over time (not surprising for an ever-changing coastline), the inability to find the appropriate dataset, incomplete or unclear restrictions on use, software and hardware incompatibility, and integration problems with other data resources (Wright, 2009).

This case illustrates some of the many technical, policy, organizational, and other considerations in the creation, use, and reuse of scientific data, but yet others also exist. For example, a great deal of anxiety persists in the larger scientific community that researchers will not be adequately recognized, that data will be reused improperly, and that data quality in general may suffer (Borgman, 2007; Gardner et al., 2003). Other concerns include worries about privacy protection for research subjects that may be identifiable in specific datasets, licensing agreements, and the legal and policy frameworks within which scientific data sharing occurs (Eckersley, Egan, De Schutter, Yiyuan, Novak, Sebesta, et al., 2003). These concerns supplement many of the purely technical challenges, such as the development of controlled vocabularies and usable interfaces for simple access and ingest.

Neuroscience is a discipline in which community-level data sharing has been successful due to standardization. Palmer, Cragin, and Hogan (2004) note that neuroscience is highly reliant on the integration of data—including models, clinical data, and basic science data—across scales and research products. One initiative, entitled Collaborative Research in Computational Neuroscience (CRCNS), has been jointly funded by the National Science Foundation (NSF) and the National Institutes of Health since 2002 to facilitate the sharing of high-quality datasets in neuroscience. Researchers polled at the onset of its development overwhelmingly expressed the need for shareable datasets to advance their work. Previous data collections in the field focused narrowly on specific data types, analytical tools, and neuroimaging data (Whyte et al., 2008). CRCNS attempts to mitigate problems through a simple and sustainable administrative structure, a stripped-down metadata scheme to tag the datasets, peer validation of datasets, and levels of access control. Subsequent funding is being made available for further projects that will contribute data to CRCNS. Another community-level collection through which neuroscience data can be shared is a visual one—an atlas of the brain that is connected to other kinds of information (Boline, Lee, & Toga, 2008). Semantic linking and spatial registration link key information to areas of the brain mapped in the atlas.

The Visible Human Project is another such collection with similar goals—the construction of a suite of minutely detailed representations of the human form. The datasets serve as anatomical references to test medical imaging algorithms; the data have also been used for a much wider range of scientific and other applications (including virtual reality and digital art). The long-term goal is to permit visual knowledge of the human body to be connected with what the National Library of Medicine (2009, online) calls "symbolic knowledge" (such as the names of body parts). Other open access medical image repositories are similarly useful for purposes ranging from regulatory review to the development of new methods for data analysis (Vannier, Staab, & Clarke, 2003).

Community-level collections represent an intermediate scale and scope between the often ad hoc self-archiving initiatives of individual research groups and large-scale reference collections. These collections face challenges in obtaining sustainable funding and other resources; implementation of standards; and the integration of multiple, often informally managed, datasets from many researchers.

Reference Collections

Reference collections, the broadly scoped, widely disseminated, well funded collections that support the research needs of many disciplines, enable data sharing in a highly formalized technical infrastructure. The underlying technologies are generally opaque with a sophisticated set of tools with structured processes designed for ingestion, normalization, and validation of both the data and metadata. These reference collection

infrastructures provide additional tools for data access, such as searching, browsing, and downloading. Several of these collections also provide analysis tools and data readers either to download material for use on the researcher's own computer or for use within the collection's infrastructure.

The oldest of these reference collections is the Worldwide Protein Data Bank (wwPDB), the international archive for biomacromolecular structural data (Berman, Bhat, Bourne, Feng, Gilliland, Weissig, et al., 2000; Berman, Henrick, & Nakamura, 2003). The wwPDB has a number of supporting organizations including the Research Collaboratory for Structural Bioinformatics (RCSB) PDB (U.S.), PDBe (Europe), and PDBj (Japan) (Berman, Henrick, & Nakamura, 2003). The wwPDB is typical of large reference collections and serves as an exemplar. Originally developed in 1971 at the Brookhaven National Laboratories as an archive for biological macromolecular crystal structures, it grew from seven to 59,000 structures (as of June 9, 2009) through improved data gathering technologies, such as nuclear magnetic resonance (NMR), and through changing research norms. Specifically, changes in the community's views about data sharing, many journals' requirement to provide a wwPDB accession code for publication, and funding agencies' requirement of data deposition for all structures led to the rapid growth of the data in the wwPDB (Berman, Westbrook, Feng, Gilliland, Bhat, Weissig, et al., 2002; Westbrook, Feng, Chen, Yang, & Berman, 2003).

With the primary goal of open access to data, a reference collection such as the wwPDB needs to have a nuanced approach to data security that balances open access and data integrity. Search, discovery, and data access are completely open. Anyone with an internet connection and a web browser can find protein structures; read the research papers; and download the data, metadata, documentation, and data analysis tools. The wwPDB has a login function but, rather than restricting or securing data or functions, it allows users to execute personalized queries either on command or on a schedule. The personalized data are stored in a secure server and not shared with third parties. No authorization or authentication is required to deposit data into the wwPDB. In addition to the automated tools that verify technical accuracy and compatibility, the wwPDB embargos new deposits until a peer-reviewed paper is published, thus ensuring intellectual verification. The wwPDB regularly reviews the embargoed data and notifies the depositors of the status of their data (Research Collaboratory for Structural Bioinformatics Protein Data Bank, 2009).

Reference collections such as the wwPDB have become an integrated component of the scientific workflow for entire domains by providing significant incentives for both the deposit of original data and the reuse of existing data. These reference collections have acquired commitments for long-term funding; they have invested in organizational and management infrastructure to ensure the persistence of the service and the data.

Technical Infrastructure

Repositories

For many data collections, a data repository is a significant component of the technical infrastructure. Through much of the literature on the technology infrastructure of scientific data sharing, the term *repository* refers to a simple data store for datasets (Venugopal, Buyya, & Ramamohanarao, 2006). We take a broader view and define a repository as a system and set of services designed as an archive for digital data with context, fixity, and persistence. Moore and colleagues (2000a; 2000b, online) describe such a digital repository as a "persistent archive" that is modular, allowing any component to be replaced without affecting the rest of the system. The low-level functional components can be updated as technology evolves so the archive remains stable in the context of rapid change. Persistent archives provide repository services to ensure the long-term archiving of and continuous access to data including backups, contingency planning, process resumption planning, hardware and network redundancy, automatic failover, and site mirroring (Kowalczyk, 2007). Repository services are mindful of the entire data life cycle (Hank & Davidson, 2009) and increase in importance as the amount of data grows (Choudhary, Kandemir, No, Memik, Shen, Liao, et al., 2000).

Ensuring that a repository is trustworthy has been a significant concern of the digital preservation community from its inception. In a seminal work, Waters and Garrett (1996) proposed an audit and certification process to evaluate repository services. Multiple initiatives have launched criteria and metrics for repository audits as methods of control and assessment (Day, 2008). In the U.S., the Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC) is a set of digital preservation best practice criteria used to review the organizational infrastructure, digital object management, technologies, technical infrastructure, and security of a repository service (Center for Research Libraries & Online Computer Library Center, Inc., 2007). The TRAC criteria are based on four principles: documentation as evidence for audits; transparency in practice, design, and policy; adequacy of implementation decisions to ensure that the preservation objectives can be met; and measurability to provide objective goals for evaluation. Several European initiatives have been undertaken in this arena and some have international import. For example, Germany has several certification initiatives including the Network of Expertise in Long-term Storage of Digital Resources (NESTOR) that evaluates the technical, organizational, and financial characteristics of a digital repository (Dobratz & Neuroth, 2004) and the German Initiative for Network Information criteria that uses visibility of the services, policy, security, authenticity, data integrity, and metadata as evaluation criteria (Deutsche Initiative für Netzwerkinformation, 2007). Ross and McHugh (2006) argue that much of the effort in repository audit and certification has been to define

the characteristics of a trusted repository; insufficient effort has been spent on developing metrics for evaluating a repository within its context and environment. They propose an evidence-based evaluation process using documentary evidence, observation of practice evidence, or testimonial evidence.

In general, specific technology infrastructures for data collections are difficult to discern from the data sharing literature. The repository services provided by data collections are not specifically addressed. The repository system software underlying data collections is generally opaque. But a number of digital repository systems exist that are used or could be used for data collections. The three major repository systems, DSpace, Fedora, and iRODS, are all open source projects governed by nonprofit foundations. Both DSpace and Fedora were developed within the library community and are jointly governed by DuraSpace. iRODS was developed in the cyberinfrastructure, grid computing, and supercomputing communities and is governed by the Data Intensive Cyberinfrastructure Foundation (Rajasekar, Wan, Moore, & Schroeder, 2006).

DSpace

Originally a joint project of the Massachusetts Institute of Technology (MIT) and Hewlett-Packard, DSpace is an open source repository system for the dissemination of digital research and educational materials (Smith, Barton, Bass, Branschofsky, McClellan, Stuve, et al., 2003). DSpace is a three-tiered system with an application layer, a business logic layer, and a storage layer (Massachusetts Institute of Technology, 2004). The layers have a strict hierarchy. Each layer has its own Application Programming Interface (API). Authentication and authorization are controlled at the application layer. DSpace is specifically designed to allow different communities within an organization to define their "space" with a set of depositors and an intellectual organization of the contents. Access to the content is through a customizable web portal.

DSpace is often implemented as an institutional repository. As a data repository, it provides a simple interface for both context data entry by the data producer and for researchers looking for data but, because it supports only Dublin Core as a metadata format, DSpace has limited ability to support robust data sharing. However, several projects use DSpace as their repository platform, including the DataShare project at the University of Edinburgh and the U.S. National Evolutionary Synthesis Center's Dryad project (DSpace, 2010).

Fedora

Fedora—the Flexible and Extensible Digital Object Repository Architecture—was originally developed as a research project at Cornell University's Computer Science Department with funding by the U.S. Department of Defense Advanced Research Projects Agency (DARPA)

and NSF, and was further developed by the University of Virginia Library and Cornell University through a grant from the Andrew W. Mellon Foundation (Lagoze, Payette, Shin, & Wilper, 2005). The major functions of Fedora are Extensible Markup Language (XML) submission and storage where digital objects are stored as XML-encoded files that conform to an extension of the Metadata Encoding and Transmission Standard (METS) schema. Fedora is the underlying repository system for a number of major digital libraries as well as several data-centric applications, such as the eSciDoc project for the research results and materials of the Max-Planck Society and the Islandora project to create a collaboration and digital data stewardship environment (Fedora Commons, n.d.).

iRods

The Integrated Rule-Oriented Data System (iRODS) is a data grid repository system developed at the University of California San Diego to support data grids, digital libraries, and persistent archives (Data Intensive Cyber Environments, 2009). As its name suggests, iRODS is a rules-based system, enforcing management policies that are sets of assertions about the digital collections. The iRODS Rule Engine interprets the management policies that have been expressed as rules determining the appropriate responses to requests, events, and conditions (Rajasekar et al., 2006). iRODS can function as a stand-alone system or can be integrated with other repository systems such as DSpace and Fedora to provide additional preservation functions; because of the rules-based processing, preservation functions such as format migration and fixity validation can be automated (Moore & Smith, 2007). iRODS is the underlying repository system for the production environments for the Teragrid and the Southern California Earthquake Center and for the National Archives and Records Administration's (NARA) transcontinental persistent archive prototype (TPAP) project.

Data Storage

For data repositories, storage infrastructure is an important component of data archiving and preservation. Although storage has become an increasingly cheaper commodity (Association of Research Libraries, 2006; Atkins, 2003; Hacker & Wheeler, 2007), carefully designing an infrastructure is important (Brown, 2003; Rajasekar et al., 1999). Moore and colleagues (2000a, 2000b) contend that a persistent archive needs a scalable storage infrastructure. Creating an expandable and extendable storage architecture based on new but proven technologies is the most efficient and likely the least expensive approach (Moore et al., 2000a, 2000b; Morris & Truskowski, 2003).

Criteria to assess scalability, expandability, and extensibility of storage technologies should include longevity, capacity, viability, obsolescence, cost, and susceptibility. The criteria can be used to develop a

matrix to measure media usability both in life span and in technical relevance, the amount of data that can be stored, data safety in terms of environmental trauma and data errors, and the total costs of implementing and maintaining the technologies (Brown, 2003).

Infrastructure Models

Data collections have to deal with four types of heterogeneity as defined by Koutrika (2005): system heterogeneity—different hardware platforms and operating systems; syntactic heterogeneity—different protocols and encodings within the system; structural heterogeneity—multiple data models and schemas; and semantic heterogeneity—multiple and conflicting name spaces giving different meanings to the same data and metadata. These multiple heterogeneities result in a recurrent struggle to find a balance between open access and security, standards and flexibility, rich metadata, and ease of submission. This heterogeneity necessitates complex tools to simplify the process of data sharing by facilitating interactions between data repositories (Ruusalepp, 2008). The federal government initiated development of large-scale infrastructures to deal with the heterogeneity of scientific data during the 1990s (Zimmerman, 2003). Much of the early infrastructure provided universal access to the raw data generated by "big science" instruments; the models that were developed have evolved and now support data generated by a wide variety of methods. Throughout the literature, these infrastructures are discussed in the abstract.

Centralized Infrastructure

In the centralized infrastructure model, technology and support are managed in a single organization. All services—data discovery, access, context, integrity, security, and data management—are delivered through this centralized infrastructure. The centralized model has a number of significant benefits including standardization, economies of scale, and reliability. It provides a standard platform for discovery, access, and context that is knowable by the designated community; simply put, there is a commonly understood and reliably consistent way of finding and accessing the data. A centralized infrastructure can also provide economies of scale. Hardware, particularly storage, has a lower cost of initial purchase in large quantities. Data loading processes can be automated, customized, and optimized for large-scale ingestion, which can reduce costs overall. A centralized infrastructure has the potential for exceptional reliability. With control from end to end, the infrastructure can be developed with redundancy at every level, with a high degree of normalization, and with smooth transitions between software components, which all contribute to a consistent and reliable service.

The centralized infrastructure, however, also has significant drawbacks. Unless there is a robust disaster recovery and business resumption plan, the centralized infrastructure is a single point of failure; that

is, if a critical piece of the infrastructure fails, the service would be unavailable to researchers. A centralized infrastructure can also be seen as a monopoly that controls all aspects of the data. Reduced diversity of input is also possible: A few individuals could have disproportionate influence over the service interface and functions. Although all shared data collections require community buy-in, with a centralized infrastructure it is even more vital to serve the needs of the research community. The alternative is massive failure. Ongoing funding can be a major issue for centralized infrastructure. Although economies of scale often exist, a centralized infrastructure is generally very expensive to build and maintain. With a large budget to maintain, it can be difficult to find funding agencies willing to provide ongoing support in the absence of significant new and groundbreaking research.

Decentralized Infrastructure

In the decentralized, or distributed, infrastructure model, the technology and support are independently managed by multiple organizations. In this model, each organization stores a set of datasets and provides services. These datasets and services may be overlapping or unique. The decentralized model implies a diversity of technologies and implementation strategies providing flexibility and creativity in both software development and end-user functionality. This model allows for multiple interfaces with different functionalities, permitting researchers to choose the best system for their work; such freedom of choice, however, can be more confusing than liberating. In this model, users can become confused by the different interactions as well as the different data available in each interface. A decentralized infrastructure introduces the potential for lack of interoperability. Effort must be expended to develop standards to allow data to be exchanged and merged.

Geographic dispersion can result in multiple redundancies; that is, if one site and its data are offline, other sites with their data can still be available. A decentralized infrastructure also implies duplication of effort in hardware implementation, software development, and management, which can increase operational costs. Ongoing support in a decentralized infrastructure is a complex puzzle. Unlike the large, integrated, centralized approach, a decentralized infrastructure generally has many smaller components. With these smaller components, it may be easier to find funding in a piecemeal manner to replace or enhance a specific component. Securing large-scale funding can often mean sharing among a large number of partners, with each receiving a small amount.

Federated Infrastructure

Federated infrastructure was first conceived in 1985 as "an approach to the coordinated sharing and interchange of computerized information" implemented with a set of distributed components, minimizing central authority (Heimbigner & McLeod, 1985, p. 253). The federated

infrastructure model is a multi-tiered environment with a central authority and independent yet interconnected partners. This implies an established distribution of power in which the central authority has a clearly defined set of responsibilities and control, documented in a binding agreement. The central authority sets standards, resolves conflicts, and provides resources (Semantic Counteroperability Community of Practice, 2008; Weill & Ross, 2004). A federated infrastructure provides a seamless interface to multiple, heterogeneous, and noncontiguous data sources. This implies a technology infrastructure that can query disparate databases, integrate results sets, and manage authentication and authorization across multiple independent organizations. Federations are complex, both organizationally and technologically.

A federated infrastructure can be implemented in a wide variety of configurations. Rajasekar, Wan, Moore, and Schroeder (2004) have developed a set of ten possible federation infrastructure models. The *occasional interchange* is the model with the greatest degree of autonomy for each partner, using the federation as an infrequent resource. In the *replicated catalog* model, each partner in the federation has an identical copy of the metadata catalog and the data; this allows for a high degree of fault tolerance. The *resource interaction* model allows a set of resources such as data or metadata to be replicated between partners. In the *replicated data zones* model, a subset of the partners (two or more) replicate data between themselves to form subnets of redundancy. The *master-slave zones* model has a central site, the master, at which new data are created and then replicated to the partner (slave) sites. This model provides a high degree of data security but limits the autonomy of the partners. The *snow-flake zones* model is a slight variation on the master-slave model, using a tiered replication: The master site replicates data to a set of slave sites, which in turn replicate to another set of slaves. The *user and data replica zone* model is a slight variation on the replicated data zones model, adding user data to the data that are replicated across subsets of the federation. The *nomadic zones* and the *free-floating zones* are very similar models involving individual, independent partners with asynchronous copies of the data and metadata for use by researchers in the field who cannot be online. The Nomadic zone implies no interconnectivity at all; the Free-floating zone implies occasional connectivity and data exchange. The *archival zone* model implies an offline site for backup purposes only (Rajasekar et al., 2004). The differences among these models include the degree of autonomy of each site, constraints on cross-registration of users, degree of replication of data, and degree of synchronization (Venugopal et al., 2006).

The federated infrastructure model facilitates powerful collaborations that can provide simple access to a wide array of data for researchers. But these collaborations take effort to build and nurture. Care is needed to maintain the balance of power among partners to keep the collaboration and partnership going. Implied in the federated model is cooperative software development. Coordinating this type of development effort

across partners is challenging but can bring a sense of ownership to the partners. The federated infrastructure model is highly configurable and thus is often complex and difficult to understand and support. Finding problems, primarily with access or data synchronicity issues, can be costly both in resources and time.

Cyber/Cloud Infrastructure

Cyberinfrastructure (or e-science in the U.K.) describes a suite of internet-based services around scientific information (Atkins, 2003) but not a specific technology or implementation. Both the grid and the cloud are strongly associated with cyberinfrastructure (Arms, Calimlim, & Walle, 2009; Berman, Fox, & Hey, 2003; Venugopal et al., 2006). Buyya, Yeo, and Venugopal (2008) describe the grid as a set of resources—such as supercomputers, storage systems, data sources, and specialized devices—that may be geographically dispersed and owned by multiple organizations used to solve large-scale, computing-intensive problems. By integrating resources, a grid creates a "virtual platform for information" (F. Berman et al., 2003, p. 9). This platform is used by a number of cyberinfrastrucure projects that have data sharing as a core mission, including Linked Environments for Atmospheric Discovery (LEAD), the Network Workbench, and the Social Informatics Data-Grid (SID-Grid). Data grids are designed specifically for distributed data-intensive applications that create massive datasets by providing shared and distributed data collections and specialized services for data integrity and security (Venugopal et al., 2006). The Resource Storage Broker (Baru, Moore, Rajasekar, & Wan, 1998), iRODS (developed at the University of California San Diego [Moore & Smith, 2007]), and the NERC DataGrid (Hey & Trefethen, 2002) are some of the major initiatives in the area of data grids for e-science research. The term is meant to include the human resources needed to make the technical infrastructure function, but the discourse suggests a technologically deterministic vision for web services in science (Jankowski, 2007).

The cloud is defined as a distributed system of networked computers using virtualization that allows for real-time provisioning to create the appearance of a single computing resource when used. The amount of computing power and/or storage provided is defined by service level agreements between the service provider and each customer (Buyya et al., 2008). Although this may seem to be just another cluster of grids, the model is very different. In cloud technology, virtual nodes are created on demand with a specific quality of service agreement that is negotiated via Web 2.0 interfaces. Buyya and colleagues contend that cloud services are a transformative technology. Cloud technology provides a "shared storage system than can provide powerful abstractions for convenient, efficient, and large-scale inter-service data sharing" (Geambasu, Gribble, & Levy, 2009, p. 1). Yet, it is not without challenges. Geambasu and colleagues identify three major challenges for large-scale data sharing using the cloud: creating a generalized and flexible data sharing

abstraction, resolving file/element naming and data protection issues, and resource allocation. The cloud provides technology as a service; the two major services are Software-as-a-Service (SaaS) and Infrastructure-as-a-Service (IaaS) both of which can simplify data sharing. Software-as-a-Service provides access to a suite of software tools, either as a monthly or per-use cost for commercial software, or at no cost for non-commercial software. This enables a computing environment with no capital investment, few set-up costs, and no maintenance. Infrastructure-as-a-Service provides on-demand computing cycles and data storage via a virtual machine that can be configured to execute requests. The impact of the cloud is not fully understood and the future is unclear. Although Geambasu and colleagues argue that the cloud infrastructure will eliminate the need for traditional data centers within organizations, Arms and colleagues (2009) argue that local systems operations help to diffuse expertise across an organization and thus will continue to be valuable.

Access and Discovery

Metadata, Ontologies, and Vocabularies

Access to and discovery of data resources require sufficient, targeted description of the resources themselves. This description is generally accomplished with metadata, ontologies, and vocabularies. The words metadata and ontologies have been used in multiple ways, but it is worth defining them briefly for our purposes. Ontologies are hierarchical, formal representations of the objects that constitute a specific area of interest (Gruber, 1995). In general, metadata has been used in information science to mean information about a particular representation of data (a document or set of documents, a dataset, images, or similar digital objects) (Duval, Hodgins, Sutton, & Weibel, 2002). A vocabulary consists of the properties that characterize the particular metadataset. For example, Dublin Core, which may be the most common metadata vocabulary, is applied to data to describe them for resource discovery and content management. It includes thirteen properties—such as title, author, and publication data—that characterize a particular digital resource.

Metadata can be informal or maintained and created in a more structured way as part of the dataset itself. Metadata can be as straightforward as the date a particular dataset was generated, the creators/custodians of the data, and the experimental or observational procedures that generated the data, to much more complex information about potential uses, geographic location of the data, its quality, or its state of completeness. Some metadata are also used to describe the kinds of web services that generated and maintain the dataset, associated tasks, and previous uses by others. Metadata can be human-generated or automatically generated and can be intended either for human use or for use by automated processes. Scientific communities, national scientific bodies, and international scientific organizations have all created metadata

standards for datasets; their complexity and use depend extensively upon the expense and resource-intensiveness of the data collection procedures. The integration of standards is necessary to make workflows and resources reliable and discoverable. Lack of metadata is cited as one of the major barriers to data sharing because it affects both machine-readable processes and human searching (Wright, 2009).

Ontologies are more complex because they purport to offer a closer match to the real world than do metadata. Ontologies offer a structured view by representing more information in the form of types, properties, and relationships. The terms that actually constitute the ontology may vary widely, but the purpose of ontologies is to model concepts for a specific domain.

A key issue with metadata and ontologies, with respect to data sharing, is the sheer diversity of both and the lack of common vocabulary among the different systems, which hinders compatibility. How should we differentiate among types of metadata? One typology of metadata from the digital library field comprises administrative, structural, and descriptive metadata. Administrative metadata describe the data necessary to manage the digital object and include such data as terms and conditions of use, provenance, and technical data (file format, data schema, etc.). Structural metadata make the object usable, such as file sequencing and time-based data and file interrelationships. Descriptive metadata include data about the intellectual content, such as creator name, title, and description (Schwartz, 2000). Rajasekar and Moore (2008) developed a layered architecture to characterize metadata: 1) kernel metadata is the most basic metadata that includes information about how the metadata is organized, information about catalogs, information about how two sets of metadata can be interoperated, and so on; 2) system-centric metadata contain systemic information including data replicas, location, size, ownership, access control information, the use of resources (computation platforms used for deriving the data, methods used, etc.), and information about users and storage resources; 3) standardized metadata include metadata that implement agreed-upon generalized standards (such as Dublin Core, MARC [MAchine-Readable Cataloging, a standard data format for bibliographic information exchange], and METS [a standard for encoding metadata in digital libraries]); 4) domain-centric metadata include metadata that implement agreed upon standards inside (and possibly across) scientific disciplines such as FITS (Flexible Image Transport System, a data standard for the transfer of astronomical data), FGDC (Federal Geographic Data Committee, pertains to the transfer of geospatial metadata), and AIML (Artificial Intelligence Markup Language); 5) application-level metadata include information that is particular to an application, experiment, group, or individual researcher and may include creation or derivation notes by individual researchers.

Provenance is an important category of metadata that tracks the steps used to derive the data from their origins to their current state

(Buneman, Khanna, & Tan, 2000). Also known as lineage, provenance is an audit trail from data creation to final state, a record of each migration, transformation, and/or process applied to a set of data (Pearson, 2002). Provenance data have no metadata standard and have been represented as graphs, semantic networks, and entity relationships (Bose, 2002). In data sharing, whether generalized or domain specific, data provenance is difficult to collect, convey, and preserve. Defining the data required to prove provenance is specific to each data type and/or domain. Although much of the research is focused on describing data for sharing, it is also an important component to preserving data. Understanding the full history of a dataset and describing its transformations is a primary function of a digital archive. Without provenance data, objects cannot be considered preserved (Chen, 2004).

Developing appropriate and effective ontologies is an essential but often thankless task (Saltz, 2002). Domain specific provenance schemas are being developed—SNOMED and LOINC for medical research (Saltz, 2002) and Karma for atmospheric research (Plale, Ramachandran, & Tanner, 2006; Simmhan, Plale, & Gannon, 2005, 2006) are cases in point. But a more generalized solution is needed. Self-describing data and the software to process those data are not yet realities (Bose & Frew, 2005). Chimera, a prototype generalized data provenance system, has attempted to abstract data representation by using types, descriptors, and transforms. Types provide an abstraction layer for semantic content, the format of the physical representation and the format's encoding; descriptors define an interpretable schema that defines the data—number of files, types of files, and so forth; and forms are a generic, typed abstraction for computational procedures (Foster, Vockler, Wilde, & Zhao, 2002). Although many research efforts have focused on acquiring and presenting provenance data seamlessly, how such metadata are applied and used in practice and how the presence of structured and searchable metadata influences and facilitates data sharing is still under investigation (Niu & Hedstrom, 2008).

One highly successful case of scientific metadata and their uses is the Flexible Image Transport System. First developed in the 1970s as a format for file exchange of astronomical data, it became widely used as a machine-readable format for online data that can be read and analyzed by data analysis software. FITS also contains a human-readable ASCII component that allows users to investigate a file of particular interest fairly readily. FITS is endorsed by NASA and the International Astronomical Union. A working group at NASA is responsible for regular updates to the content and structure of FITS.

Data Access

Within the context of data sharing, data access is defined as the process by which a researcher can obtain data in order to act upon them. The most challenging technical issue is actually how to deliver large, often

petabyte, datasets to researchers over networks with insufficient bandwidth (Gray, Liu, Nieto-Santisteban, Szalay, DeWitt, & Heber, 2005). Lubell, Rachuri, and Mani (2008) define a taxonomy of access scenarios for data collections—reference, reuse, and rationale. Reference is the ability to read the digital object; reuse is the ability to modify the digital object in an appropriate system environment; and rationale, the highest level of access, is the ability to explain any decisions that are made about the content of the digital object. In order for researchers to be able to reference and reuse data, they must have access.

Three models of data access have emerged: moving the data to the tools; moving the tools to the data; and building virtual views and dynamic access. Moving the data to the tools was the first model. Before organized data collections developed, data were shared by moving the data from the data owner's computer to another researcher's computer via email, FTP, Network File System (NFS), and other simple technologies (Birnholtz & Bietz, 2003; Gardner et al., 2003). As early data collections developed, this first model was prevalent. It was simple to implement and allowed the researchers to use their preferred tools within their own computing environments. Over time, both the size of datasets and the use of the collections grew, resulting in longer download times and increased demands on the collections' network infrastructure (Arms et al., 2009).

With the expansion of data collections, a second model has developed. Rather than moving the data to the researcher, the software tools are moved to the data (Arms et al., 2009; Gray et al., 2005; Michener et al., 2005; Myers et al., 2005). The tools can be specific to a data format, such as mesoscale weather analysis (Droegemeier, Gannon, Reed, Plale, Alameda, Baltzer, et al., 2005), or specific to a function, such as spatial mapping (Michener et al., 2005). The number of tools is proliferating, raising yet another discovery issue: How do researchers find the right tool for their needs? Many data collections provide a discovery portal for the tools that they offer (Myers et al., 2005). Providing a generalized infrastructure for tool delivery is a new research stream with emphasis on data conversion and integration (Myers et al., 2005; Schuchardt, Pancerella, Rahn, Didier, Kodeboyina, Leahy, et al., 2007).

A third model is evolving: data integration. This consists of providing the infrastructure to build new, dynamic data resources on demand by creating virtual views of slices of different data collections. Two methods are emerging; one uses schema mapping and the other semantic annotations and ontologies. The schema mapping method focuses on replication, synchronicity, and constraint issues of dynamic data exchange between collaborating researchers. Arenas, Kantere, Kementsietsidis, Kiringa, Miller, Mylopoulos, and colleagues (2003, p. 54) refer to mappings between seemingly unconnected databases as "mediation across multiple worlds" that can apply rules to check automatically for data consistencies and infer new constraints. Ives, Khandelwal, Kapur, and Cakir (2005) suggest that trust policies could specify the conditions

under which a database is shareable and could score the incoming data based on perceived quality or relevance. These scores would then be used to reconcile conflicts. Like other rules-based processing, schema mapping requires a significant investment to build the necessary rules. The second method for an infrastructure for data integration uses semantics mediation to link datasets to ontologies explicitly. The semantic annotations are used to drive a knowledge-based data integration service that performs a series of integration transformations, such as scale resolution and geographic projection reconciliation, as well as data mining services, such as sampling data to create vectors with layers associated with an occurrence point. Ontologies built for data integration can have multiple purposes, such as thematic searching and metadata creation (Michener, 2005). Developing automated tools to build the ontologies and generate variable-names metadata in datasets that lack attribute-level metadata is ongoing (Jaiswal et al., 2006).

Metadata and ontologies function as identification for data; but sustainable access to data requires a method for providing persistent identification (Helly, Elvins, Sutton, & Martinez, 1999; Helly, Elvins, Sutton, Martinez, Miller, Pickett, et al., 2002; Lubell et al., 2008; Rajasekar et al., 2004). That is, sustainable access requires a unique, permanent, location-independent identifier for a network-accessible resource, which, using the language of the World Wide Web, is commonly known as a Universal Resource Identifer (URI) (World Wide Web Consortium, 2001). The URI is a generalized concept that is implemented in a number of different schemes: Persistent URLs, Digital Object Identifiers, and many others. In general, a Universal Resource Locator (URL) cannot be a persistent identifier because it conflates two important but separate functions—item location and item identification. By separating location from identification, URIs help avoid the problem of broken links. URIs are generally implemented as URLs that require a multiphased resolution process. Figure 6.1 shows the generalized model of this process (Harvard University Library, 2003). The first level resolution is to find the right resolution server. Once the resolution server is located, the second level resolution finds the resource. The second phase of the resolution is relatively simple. The resolution service uses a database that stores both the persistent ID and a location of the object (usually in the form of a URL) to redirect the http transaction using the location URL. As the problem of persistent identification of digital resources has become more prominent, a number of competing URI schemes have been developed.

The Online Computer Library Center (OCLC) uses a PURL—a persistent URL. This technology is similar to the Harvard persistent identification system but with a different syntax (Shafer, Weibel, & Jul, 2001). With funding from DARPA, the Corporation for National Research Initiatives (CNRI) created another unique persistent identifier syntax, the Handle System. The Handle System is an open source, downloadable technology specification with a set of open protocols and a

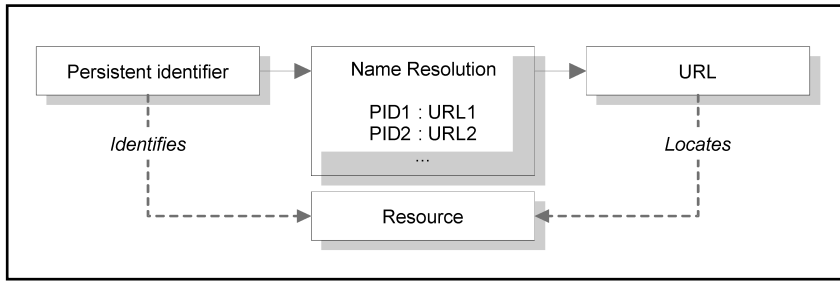

Figure 6.1 Harvard University's Name Resolution Service.

namespace (www.handle.net). The DSpace digital repository system uses Handle to identify its contents uniquely and persistently. The International DOI Foundation developed the semantics for the Digital Object Identifier (DOI), which has been implemented using the Handle syntax and resolution system (www.doi.org). Although the DOI specification was originally developed by publishers of electronic information to provide persistent and controlled access to journal articles, DOIs are often used to cite data in experiments and in provenance schemas (Simmhan, Plale, & Gannon, 2005). The California Digital Library has developed the Archival Resource Key (ARK), which provides access to a digital information resource. The ARK is architected to deliver three parameterized services: the digital object metadata, digital object content files, and a commitment statement made by the owning organization concerning the digital object (Kunze, 2003).

Data Discovery

Data discovery needs for data collections are both general and specific. The problem is "finding both needles in the haystack and finding very small haystacks" (Gray et al., 2005, p. 34). Researchers need a query interface that allows for a general overview of the complete collection to understand the breadth and depth of the data available. Beyond the general view, the data discovery system within a data collection commonly allows queries by subject terms, by data creator, and by data format, as well as by a more general set of terms (Treloar & Wilkinson, 2008). Domain specific ontologies are beginning to be used for concept-based searching (Jaiswal et al., 2006; Michener, 2005). Other types of discovery are emerging, such as geographic and spatially oriented searches and navigating provenance trees (Myers et al., 2005). Although researchers have indicated an interest in discovering data across domains, data sharing that crosses such boundaries is inhibited by the innate differences in the information-seeking behaviors of researchers within those different domains as well as the disparate vocabularies used (Kingsley, 2008; Treloar & Wilkinson, 2008). Parsons and Duerr

(2005) contend that data discovery interfaces require multiple views based on the designated user community in order to minimize unanticipated, inappropriate use of data.

Many of these data collections provide discovery services through a portal (Ruusalepp, 2008). A metadata portal is an internet site that provides an integrated access interface to distributed information resources (Myers et al., 2005; Schindler & Diepenbroek, 2008; Wright, 2009). Portals are also used for cross-domain searching (Treloar & Wilkinson, 2008). Portals can reduce the complexity of the underlying infrastructure for researchers. In distributed search infrastructures, every data provider has a separate metadata catalog and search interface; a domain portal can create a unified interface by federating searches and integrating result sets (Schindler & Diepenbroek, 2008). Although some projects use the library-oriented, client-server protocol Z39.50 to support federated searching (Altman, Andreev, Diggory, King, Sone, Verba, et al., 2001; Schindler & Diepenbroek, 2008), the U.K.'s Joint Information Systems Committee (JISC) recommended the A9 OpenSearch specification, a Web 2.0 technology for sharing search results as syndications and aggregations, in its most recent report on portal development in e-science (Allan, Crouchley, & Ingram, 2009).

Portals used to harvest data from a wide variety of sources, and then integrate those beyond the original scope of the data collection, are emerging (Ruusalepp, 2008). The primary technology for this harvesting is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), a simple, six-step protocol for metadata exchange (Duke, Day, Heery, Carr, & Coles, 2005; Jaiswal et al., 2006; Schindler & Diepenbroek, 2008). As a simple protocol, OAI-PMH is difficult to implement in a multifaceted data environment because Dublin Core is the only natively supported metadata format, creating difficult choices when mapping complex data streams. Consider the following OAI-PMH data mapping example from a crystallography data collection project:

> [The] duality of the role of data contained within the data files, which at times was part of the data, but at other times fulfilled the role of exchanged metadata, or simply additional information displayed to a user, made discussion and reconciliation of views difficult, particularly at the initial stages when a common ground of understanding between the different partners was still being established. (Duke et al., 2005, p. 52)

To resolve the issues of mapping complex data into a simple data structure, the Open Archive Initiative has developed a new data representation protocol specifically for complex data objects, the Object Exchange and Reuse, commonly referred to as OAI-ORE (Lagoze & Van de Sompel, 2007). The OAI-ORE was designed to make the information within a complex object discoverable, reusable, machine readable, and

actionable (Cheung, Lashtabeg, Drennan, & Hunter, 2008). Complex data objects are composite objects; that is, they consist of multiple independent data objects such as a binary dataset, a set of TIFF images, and a text document. The OAI-ORE protocol uses named graphs to describe the relationships among the many parts of a complex object. Each individual data object (a node) has one or more relationships (arcs) to other objects. Nodes and arcs are web resources and have URIs. Each element can be reused and recombined to make new named graphs, meaning that a dataset can be linked to multiple published and working papers with minimal duplication of data (Cheung et al., 2008). The Center for Embedded Networked Sensing (CENS) used the OAI-ORE protocol to provide interoperability for its content (Borgman, Wallis, Mayernik, & Pepe, 2007; Pepe, Borgman, Wallis, & Mayernik, 2007).

Data Security

Data security refers to the "protection of data from accidental or intentional but unauthorized modification, destruction or disclosure through the use of physical security, administrative controls, logical controls, and other safeguards to limit accessibility" (U.S. Social Security Administration, 2009, para. 10). Securing data collections may involve ensuring against corruption of content (integrity), safeguarding privacy (anonymity), or verifying users' identities (authentication) (Venugopal et al., 2006). Securing content against corruption is a matter of ensuring the fixity of the data object and is a responsibility of the data repository. Authentication—verifying the identities of those using a system or a dataset and determining the functions that the authenticated person is allowed to perform—is based on the role that a person has within the context of that data sharing system or a set of rules about the dataset itself. Personal identity authentication could be managed at the collection level, at the storage level, or at the home institution of the person: The collection could manage a control list of each person allowed to access the data (Rajasekar & Moore, 2008); a storage cloud could secure the data by requiring an encrypted signature as certificate of authorization (Geambasu et al., 2009); the home institution of an individual could authenticate the person and then pass a token to the collection as a certificate of authorization using a system such as Shibboleth (Scavo & Welch, 2007). Roles-based authorization defines a prescribed set of behaviors or functions that can be performed within a system such as resource discovery, resource acquisition, data upload, or metadata creation. These behaviors have associated polices that define the relationship between the authenticated person and the authorized behaviors (Jin & Ahn, 2006a). Within data collections, the behaviors are tied to specific data, creating a hybrid approach to authorization (Jin & Ahn, 2006b).

Human Infrastructure

We use this term to discuss many of the social, cultural, and ethical issues that arise around data sharing and how these are managed through policy and practice. Human infrastructure encompasses policy-related issues and scientific practice, although some of these ideas have been discussed elsewhere in this chapter as appropriate.

Policy

Policy on data sharing and use is a broad topic. Arzberger, Schroeder, Beaulieu, Bowker, Casey, Laaksonen, and colleagues (2004) articulate three domains that govern policy making and enforcement: the governmental, institutional, and cultural. The levels at which policy is made and enforced encompass national and international standards and agreements, institutional and other data repositories, guidelines at specific institutions where research data are generated and housed, and journal standards for publication of research data and results. In the U.S., for example, the National Institutes of Health (2003) requires grant applications of greater than $500,000 to submit a plan for data sharing. Many journals routinely require publishing researchers to submit their data to standardized repositories, complete with metadata (examples include the *Proceedings of the National Academy of Sciences* and *Molecular and Cellular Proteomics*, which has been requiring data submission since 2004). Professional societies such as the Microarray Gene Expression Data Society also promote data sharing by requiring researchers to submit their primary data to a public repository and provide a checklist to help achieve this goal. Such journals and societies have policies for ensuring privacy of the research data, including security mechanisms and de-identification of personal data.

Universities and other research institutions have developed a range of policies in the realm of data sharing in response to a variety of pressures. Institutional Review Boards (IRBs) in the U.S. and Research Ethics Boards (REBs) in the U.K. that govern human subjects research are a major influence. These entities place strictures on how data should be managed, stored, and shared to maximize benefit and minimize risk when human subjects are involved in research. These policies create incentives for data sharing but also introduce potential conflicts that inhibit it (Piwowar & Chapman, 2008).

Privacy

In the past, most data with highly personal information, primarily medical data, were simply not shared or were de-personalized by removing names and addresses (F. Berman et al., 2003). But even with de-personalized data, it is possible to identify individuals by triangulating multiple datasets. New techniques are being developed to de-indentify individuals in datasets through different automated anonymizing software tools

including population-density-based Gaussian spatial blurring (Cassa, Grannis, Overhage, & Mandl, 2006) and the k-anonymization algorithm that ensures each record is indistinguishable from the others (Bayardo & Agrawal, 2005).

Data mining tools can be used to circumvent these operations. For example, a recent study found that some genomics data and research papers could be mined to reveal the identities of genetic disease patients even though data had been de-identified in genome-wide analysis studies (Kaye, Heeney, Hawkins, de Vries, & Boddington, 2009). Anonymization and informed consent are often used to protect the identity of research subjects, but both are problematic in the data sharing context. DNA, for instance, is a unique identifier; it is almost impossible to completely de-identify DNA and doing so would potentially make the research less useful. Informed consent protocols routinely take into account the uses of de-identified data but few consider the implications of aggregation for data sharing (Karp, Carlin, Cook-Deegan, Ford, Geller, Glass, et al., 2008).

Geographical information systems (GIS) provide another case of data-intensive technologies that illustrate some of the pitfalls associated with individual and demographic identification. Although individuals are not identifiable through GIS spatial coordinates, the potential for data mining and the aggregation of data may make sensitive data more visible. In particular, with aggregated GIS data, there is potential for racial profiling, redlining, and other potentially undesirable ways of segmenting social groups.

Ownership

Few financial rewards exist for sharing information; GIS layers are a potential commodity and thus create intellectual property dilemmas. A significant amount of GIS information is created by government agencies at local, state, and federal levels, which in the context of U.S. law and policy might suggest that GIS information should be freely shared because it is funded through public monies. However, many agencies are reluctant to share public data, knowing that private entities could potentially profit from public information while not sharing their own data. It is also unclear whether GIS data count as a public good in the same manner as other state-held resources. This creates concerns about how GIS should be used for commercial purposes.

Data sharing practices in science vary greatly. Intellectual property rights and the fair use of data remain problematic concepts for the sharing of scientific data. Because datasets are not formal publications or raw facts and are copyrighted but in the public domain (if generated with federal funds), they occupy a gray area in the arena of licensing and use restrictions and have not been governed by extant intellectual property regimes. Indeed, whether datasets are facts or publications is still contested in copyright law. One model from the Massachusetts Institute

of Technology is the Science Commons, which takes a Creative Commons licensing approach to scientific data and indeed, uses the Creative Commons infrastructure. Scientists can use Science Commons to establish rights of attribution, sharing, and commercial/noncommercial use for datasets and other scientific data. Science Commons is not yet widely used or integrated into data repositories, although numerous projects exist as proof-of-concept, especially in the neurosciences.

Trust

Many factors affect the "quality" of data, which in turn influences how and when data are reused. One key element is the role of trust that researchers have in particular datasets. Extensive literature exists on the role of trust in data and its importance for data sharing, but there also has been a great deal of interest in the role of trust in the use of digital information, particularly in the context of digital libraries (Van House, 2002). The formal publication system eases some concerns through the process of peer review, as do the larger research collections with their vast resource-laden mechanisms for vetting data; trust in community-level collections, however, is not necessarily managed in the same way. For example, Van House, Butler, and Schiff (1998) describe two intermediate processes through which researchers in the environmental sciences decide whether to trust the quality of environmental datasets. First, these researchers take into account the scientific processes that were employed in creating the data. However, this condition, although necessary, is not sufficient. How data are collected and interpreted can be consciously or unconsciously biased by the creators, they argue. Thus the second criterion by which data trustworthiness is assessed is the personal and professional reputation of the individual, group, or organization that generated the dataset. Researchers often argue that they know how trustworthy data will be, based on the professional reputations of the creators (Shankar, 2007; Van House et al., 1998). These trust-building and credibility measures cannot be universally prescribed or formally built into digital data repositories; they are local, contingent, and fluid and are as much about the social conditions under which the data were generated as any qualities inherent in research design and method (Kelton, Fleischmann, & Wallace, 2008).

Local Practice Effects

The field of genomics (the characterization and sequencing of whole genomes of organisms) is one area that is widely cited as an exemplar of data sharing. Much of this has been driven by the need for large numbers of samples and declining costs and greater efficiencies of sequencing (Kaye et al., 2009). Funding agencies have made data sharing a requirement of financial support (Nelson, 2009; Schofield, Bubela, Weaver, Portilla, Brown, Hancock, et al., 2009). There are also extensive non-financial incentives for researchers to share data (Research

Information Network, 2008). Because the sequencing of a genome and its widespread acknowledgment by others generates credit for the sequencing laboratory or group, data sharing enhances reputation. Strong data description standards (with a limited and easily used metadata schema), limited data types, journal policies that require submission to repositories as a condition of publication, and a narrow scope of users (generally, other genetics researchers who are trained to use the data) further promote data sharing. Birnholtz and Bietz (2003) argue that genomics and similar disciplines where data sharing is relatively straightforward and routinized are characterized by low task uncertainty and the mutual interdependence of researchers. The relatively small community of genomics researchers agree on what the research problems are in the field, what kind of data are needed to address them, and the methods by which the data are produced; they rely on each other to provide such data. Few studies have been conducted in other disciplines, so it is not clear if this phenomenon is domain-specific.

Evaluation

Data repositories of all types are resource-intensive operations and represent substantial investments of time and money. It stands to reason that they require systematic evaluation for multiple reasons—the value of building the system, its usefulness and usability, importance to researchers, and other metrics. Some of these metrics are quantitative, others more interpretive. They represent the goals of different stakeholders in the data-sharing process. In short, evaluating the impact of data repositories and data sharing requires complex, extensive, often longitudinal study. Hilgartner and Brandt-Rauf (1994) argue that many studies of scientific practice have traditionally sidestepped data sharing as an explicit research problem.

However, this lacuna has since been addressed through numerous mixed-methods case studies of data access and use in cyberinfrastructure (Baker & Bowker, 2007; Karasti, Baker, & Halkola, 2005; Ribes & Finholt, 2007; Sonnenwald, Whitton, & Maglaughlin, 2008), scholarly communication in research collections (Palmer, 2005), and institutional repositories (Davis & Connolly, 2007). Many of these studies have been qualitative in nature and have emphasized local practices that emerge and shape the design and use of these resource-intensive infrastructures, often for the purpose of more effective design. Other recent turns include a call for more policy analysis, particularly with respect to institutional and systemic issues that influence practice (Olson, Zimmerman, & Bos, 2008). More comparative studies across disciplines and infrastructures are needed, as well as case histories of successful and unsuccessful data sharing.

The Future

Emerging technologies are being incorporated into infrastructures for sharing scientific data. Web 2.0 services are being built into generalized web services for e-science. RSS feeds (known as Really Simple Syndication or Rich Site Summary) can notify scientists of new data added to a database or of new data created by remote sensing equipment. Mash-ups, usually implemented as a client-side, data overlay service, can be more complex server-side services providing new methods of aggregating multiple sources of data into ad hoc applications (Fox, Guha, McMullen, Mustacoglu, Pierce, Topcu, et al., 2009; Geambasu et al., 2009). Web 2.0 services have been created to explore species distribution patterns, apply clustering algorithms, and integrate with GIS systems in a general workflow engine (Zhang, Altintas, Tao, Liu, Pennington, & Michener, 2006). Fox and colleagues (2009) are using Web 2.0 technologies to create small portable web services for database updates, data annotations, user management, and access rights. Infrastructure to create ubiquitous computing with transparent data exchange using wireless ad hoc peer-to-peer communications on smart phones and personal digital assistants (PDAs) is becoming widely available (Heinemann, Kangasharju, Lyardet, & Mühlhäuser, 2003). Building on these standards, systems to use ubiquitous computing infrastructure for collecting remote sensing data are being deployed (Chen & Kotz, 2001).

Web 2.0 is also being utilized in other ways, primarily by the user community. Contemporary science is influenced by competing impulses that permeate the daily work practices of scientists and the institutions in which they work. One such impulse is the need to share research results, data, and tools to promote greater impact of research, leverage scarce resources, conduct longitudinal studies, apply new analytical tools to existing datasets, and verify and refine the results of other researchers. Data sharing also has the potential to foster multidisciplinary research. The other, conflicting impulse has to do with concerns related to misuse of data, inappropriate or insufficient citation, or the desire to sequester data for the purposes of commercialization. In the scientific publication arena, these competing forces have engendered national and international discussions as well as a thriving open access movement (Willinsky, 2006). Proponents of open access argue that publishing in open access journals creates wider dissemination of results for other researchers and provides greater value to the general public, which funds most research in the U.S. Advocacy groups for greater taxpayer access to scientific research, librarians, and even national governments and the United Nations support open access. However, commercial publishers and other critics of open access have argued that there is little evidence of its merits. They contend that most people can access the literature they need, that open access would place undue burdens on researchers in some fields, and that it would not be economically viable.

Most recently, discussion has trickled down to the primary scientific data. As has been noted, although government agencies directly or indirectly sponsor a significant portion of basic scientific research, little research data is publicly accessible. Thus, the practice of science comes into conflict with its often-expressed ideals of openness, shared data, intellectual rigor, and verifiability. Advocacy groups and other stakeholders contend that open access to data promotes greater use of research data, not just by scientists but also by health care providers, individuals, and nonprofit organizations (Zuccala, 2009). However, contemporary science and its commercialization have been built upon the protection of data and records as forms of intellectual property that are closely held by the scientific community. In response to these challenges, scientists and organizations are building upon the open access movement to develop an approach to open research that respects intellectual property concerns. In the spirit of Free/Libre Open Source Software, many scientists are using portals, wikis, and blogs to share their workflows, datasets, and other preliminary research products.

A more radical approach is one in which all data and research are made publicly available from the outset:

> By [Open Notebook Science] I mean that there is a URL to a laboratory notebook that is freely available and indexed on common search engines. It does not necessarily have to look like a paper notebook but it is essential that all of the information available to the researchers to make their conclusions is equally available to the rest of the world. Basically, no insider information. (Bradley, 2006, para. 5)

This is different from open access science as discussed in the introduction, which emphasizes access to published journals through institutional repositories and other mechanisms, but builds upon a similar spirit of sharing. Although, for numerous reasons, Open Notebook Science is not yet widespread, it presents an intriguing new approach to scientific research with implications for data sharing, the development of institutional repositories, e-science and e-social science, cyberinfrastructure use, and national/international policy on data in science. Studying data sharing from the perspectives Open Notebook Science offers for collaboration, access, and ease of data management will be important to understanding current bottlenecks in access to data, verifiability, and the development and use of data repositories.

Lesk (2008) contends that a new professional field, data curation, needs to be created. Parsons and Duerr (2005) use the term data stewardship. Data curators or stewards would work with data creators to develop practices and skills to preserve data. Each scientific domain would not need to develop these in a vacuum but could use the data curator's professional skills in database design, digital forensics, and quality assurance. This new profession should have a reward system

that demonstrates the value of the field including conferences, journals, academic credentials, and standing within their organization for the data curators and academic rewards for the scientists who deposit and preserve their data (Lesk, 2008; Research Information Network, 2008).

An important arena of ongoing discussion and research on data sharing is the development of sustainable funding models for data archiving. Understandably, costs vary widely based on data and file formats, required metadata, security needs for limiting access to sensitive or private data, and ramp-up costs for initiating a new repository (Beagrie, 2006). Ingest and initial archiving tasks are generally the most expensive; the maintenance of the archive over time may become less expensive with reduction in costs of memory and computing power. Lesk (2008) contends that the preservation, persistence, longevity, and management of data are all tightly coupled with access; funding for preservation and curation activities will be based on the perceived usefulness and accessibility of the data. This is not surprising. There is usually a direct relationship between the cost of metadata creation and the benefit to the user: Describing each item is more expensive than describing collections or groups of items but clearly more useful for data discovery (Duval, Hodgins, Sutton, & Weibel, 2002). Creating sustainable funding that does not respond to short-term pressures but, instead, is predicated upon long-term data need and use is challenging (Interagency Working Group on Digital Data, 2008). Career- and grant-related incentives to produce and develop sound data management plans are part of these challenges.

The 2009 award of the Nobel Prize in Economic Sciences to Eleanor Ostrom suggests another approach to distributing costs: the model of the commons. Ostrom's work demonstrates some of the failures of market-based approaches to governing public goods and highlights successes of information commons (Hess & Ostrom, 2009). Given the complex incentive problems we have already described and a globalized intellectual property regime that complicates data reuse, the self-archiving commons model is probably insufficient. However, positive incentives for self-archiving have been demonstrated to be useful in some domains such as microbiology (Dedeurwaerdere, Berleur, Nurminen, & Impagliazzo, 2009). Incentives include greater public exposure, faster publication times, and reduction of costs for publication and access. Case studies of sustainability and cost issues show that a single approach seldom works. Instead, hybrid strategies that incorporate in-kind support from host institutions, revenue generation, volunteer and community labor, cost-control mechanisms, and a long-term commitment to keeping resources accessible are essential to a successful model. Achieving balance is not simple (Maron, Smith, & Loy, 2009).

Conclusion

This exploration clearly suggests that there are many gaps in the various literatures that constitute research and practice in data sharing and no clear way of harmonizing what is known and what is not. Both practical and theoretical questions arise from the consideration of data sharing. On a practical level, numerous stakeholders are interested in making scientific data re-usable through the development and implementation of repositories; metadata schema; standardized vocabularies; and interfaces for relatively straightforward ingest, management, and access. Still more stakeholders are concerned about the policy implications of data sharing (or lack thereof) and are concerned with barriers to and incentives for doing so, the latter often implemented through normative measures and ethical training. And yet another group considers data sharing and access as forms of exchange that underpin and illuminate the daily workings of science as well as the broader political economy of this essential form of knowledge production. And of course, there are broad overlaps among these communities of interest.

Research has been global in nature for decades, but the ease with which data can move across national boundaries, the existence of competing (and often conflicting) legal regimes around its use and ownership, and the importance of preserving scientific data for longitudinal research mark this period of scientific research. The globalization of research is being fostered by and is in turn fostering greater research and policy making on e-science/digital infrastructures, legal and policy regimes, and calls for change in scientific practice. For these reasons, it is especially critical that research and development on data sharing be at the forefront of the consciousness of those who are most affected by it and have the most to gain or lose: policymakers, technologists, the public, and scientists themselves.

Acknowledgments

We would like to thank Heather Piwowar for sharing some of her bibliographic resources with us, Shannon Oltmann for her careful editing and thoughtful comments, the anonymous reviewers for their substantial input and help, and Katy Börner for suggesting this project to us in the first place.

References

Abrams, S. A. (2004). The role of format in digital preservation. *VINE*, *34*(2), 49–55. Retrieved August 27, 2005, from www.emeraldinsight.com/10.1108/03055720410530997

Allan, R., Crouchley, R., & Ingram, C. (2009). *JISC information environment portal activity: Supporting the needs of e-research. Interim Report*. Daresbury, UK: Joint Information Steering Committee. Retrieved January 22, 2010, from epublicns03.esc.rl.ac.uk/bitstream/3692/interim.pdf

Altman, M., Andreev, L., Diggory, M., King, G., Sone, A., Verba, S., et al. (2001). A digital library for the dissemination and replication of quantitative social science research: The virtual data center. *Social Science Computer Review, 19*, 458–470.

Arenas, M., Kantere, V., Kementsietsidis, A., Kiringa, I., Miller, R. J., Mylopoulos, J., et al. (2003). The Hyperion project: From data integration to data coordination. *ACM SIGMOD Record*, *32*(3), 53–58.

Arms, W. Y., Calimlim, M., & Walle, L. (2009). E-Science in practice: Lessons from the Cornell Web Lab. *D-Lib Magazine*, *15*(5/6). Retrieved January 24, 2010, from www.dlib.org/dlib/may09/arms/05arms.html

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., et al. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, *3*, 135–152.

Association of Research Libraries. (2006). *To stand the test of time: Long-term stewardship of digital datasets in science and engineering*. Arlington, VA: The Association.

Association of Research Libraries, Association of American Universities, Coalition for Networked Information, & National Association of State Universities and Land-Grant Colleges. (2009). *The university's role in the dissemination of research and scholarship*. Washington, DC: Association of Research Libraries.

Atkins, D. (2003). *A report from the U.S. National Science Foundation Blue Ribbon Panel on Cyberinfrastructure*. Arlington, VA: National Science Foundation, Directorate for Computer and Information Science and Engineering.

Baker, K. S., & Bowker, G. C. (2007). Information ecology: Open system environment for data, memories, and knowing. *Journal of Intelligent Information Systems*, *29*(1), 127–144.

Baru, C., Moore, R., Rajasekar, A., & Wan, M. (1998, November/December). *The SDSC storage resource broker.* Paper presented at Conference of the Centre for Advanced Studies on Collaborative Research, Toronto, Ontario, Canada.

Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. *Proceedings of 21st International Conference on Data Engineering*, 217–228.

Beagrie, N. (2006). Digital curation for science, digital libraries, and individuals. *International Journal of Data Curation*, *1*(1), 3–16.

Berman, F., Fox, G., & Hey, T. (2003). The grid: Past, present, future. In F. Berman, G. Fox, & T. Hey (Eds.), *Grid computing: Making the global infrastructure a reality* (pp. 9–50). Chichester, UK: Wiley.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., et al. (2000). The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology*, *7*, 957–959.

Berman, H. M., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, *10*, 980.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2002). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Birnholtz, J., & Bietz, M. (2003, November). Data at work: Supporting sharing in science and engineering. *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, 339–348.

Boline, J., Lee, E. F., & Toga, A. W. (2008). Digital atlases as a framework for data sharing. *Frontiers in Neuroscience*, *2*(1), 100–106.

Borgman, C. L. (2007). *Scholarship in the digital age: Information, infrastructure, and the internet*. Cambridge, MA: MIT Press.

Borgman, C. L., Wallis, J. C., Mayernik, M. S., & Pepe, A. (2007). Drowning in data: Digital library architecture to support scientific use of embedded sensor networks. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 269–277.

Bose, R. (2002). A conceptual framework for composing and managing scientific data lineage. *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*. Retrieved October 10, 2009, from dx.doi.org/10.1109/SSDM.2002.1029701

Bose, R., & Frew, J. (2005). Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys*, *37*(1), 1–28.

Bradley, J.-C. (2006). *Open notebook science*. Retrieved January 23, 2010, from drexel-coas-elearning.blogspot.com/2006/09/open-notebook-science.html

Brown, A. (2003). *Digital preservation guidance note 2: Selecting storage media for long-term preservation*. Kew, England: The National Archives of England, Wales and the United Kingdom.

Buneman, P., Khanna, S., & Tan, W. C. (2000). Data provenance: Some basic issues. In *Foundations of Software Technology and Theoretical Computer Science*, *1974*, pp. 87–93.

Buyya, R., Yeo, C. S., & Venugopal, S. (2008). *Market-oriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities*. Paper presented at The 10th IEEE International Conference on High Performance Computing and Communications. Retrieved June 23, 2009, from arxiv.org/abs/0808.3558

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., et al. (2002). Data withholding in academic genetics: Evidence from a national survey. *Journal of the American Medical Association*, *287*(4), 473–480.

Cassa, C. A., Grannis, S. J., Overhage, J. M., & Mandl, K. D. (2006). A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *Journal of the American Medical Informatics Association*, *13*(2), 160–165.

Center for Research Libraries & Online Computer Library Center, Inc. (2007). *Trustworthy repositories audit and certification: Criteria and checklist*. Chicago: Center for Research Libraries.

Chaplin, M. 2004. A challenge to conservationists. *World Watch Magazine*, *17*(6). Retrieved January 22, 2010, from www.worldwatch.org/node/565

Chen, G., & Kotz, D. (2001). Solar: Towards a flexible and scalable data-fusion infrastructure for ubiquitous computing. *Proceedings of the Workshop on Application Models and Programming Tools for Ubiquitous Computing at the Third International Conference on Ubiquitous Computing*. Retrieved January 22, 2010, from cmc.cs.dartmouth.edu/papers/chen:solar.pdf

Chen, S. S. (2004). Digital preservation and workflow process. In *Digital libraries: International collaboration and cross-fertilization* (pp. 61–72). Berlin, Germany: Springer.

Cheung, K., Lashtabeg, A., Drennan, J., & Hunter, J. (2008). SCOPE: A scientific compound object publishing and editing system. *International Journal of Digital Curation*, *3*(2). Retrieved January 22, 2010, from www.ijdc.net/index.php/ijdc/article/viewFile/84/55

Chin, G., & Lansing, C. (2004). Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory. *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, 409–418.

Choudhary, A., Kandemir, M., No, J., Memik, G., Shen, X., Liao, W., et al. (2000). Data management for large-scale scientific computations in high performance distributed systems. *Cluster Computing*, *3*(1), 45–60.

Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system (OAIS), recommendation for space data system standards*. Washington, DC: The Committee.

Crow, R. (2002). *The case for institutional repositories: A SPARC position paper*. Washington, DC: Scholarly Publishing and Academic Resources Coalition.

Data Intensive Cyber Environments. (2009). *IRODS: Data grids, digital libraries, persistent archives, and real-time data systems*. Retrieved January 22, 2010, from www.irods.org

Davis, P. M., & Connolly, M. J. L. (2007). Institutional repositories: Evaluating the reasons for non-use of Cornell University's installation of DSpace. *D-Lib Magazine*, *13*(3/4). Retrieved January 24, 2010, from www.dlib.org/dlib/march07/davis/03davis.html

Day, M. (2008). Toward distributed infrastructures for digital preservation: The roles of collaboration and trust. *International Journal of Digital Curation*, *1*(3), 15–28.

Dedeurwaerdere, T., Berleur, J., Nurminen, M., & Impagliazzo, J. (2009). Databases, biological information and collective action. In J. Berleur, M. Nurminen, & J. Impagliazzo (Eds.), *Social informatics: An information society for all? In remembrance of Rob Kling* (pp. 159–169). New York: Springer.

Deutsche Initiative für Netzwerkinformation. (2007). *Electronic publishing: DINI-Certificate: document and publication services*. Retrieved January 22, 2010, from www.dini.de/english/dini-certificate

Dobratz, S., & Neuroth, H. (2004). Nestor: Network of expertise in long-term storage of digital resources: A digital preservation initiative for Germany. *D-Lib Magazine*, *10*(4). Retrieved January 24, 2010, from www.dlib.org/dlib/april04/dobratz/04dobratz.html

Droegemeier, K. K., Gannon, D., Reed, D., Plale, B., Alameda, J., Baltzer, T., et al. (2005). Service-oriented environments for dynamically interacting with mesoscale weather. *Computing in Science & Engineering*, *7*(6), 12–29.

DSpace. (2010). *Repository list*. Retrieved January 22, 2010, from www.dspace.org/whos-using-dspace/Repository-List.html

Duke, M., Day, M., Heery, R., Carr, L. A., & Coles, S. J. (2005, March). *Enhancing access to research data: The challenge of crystallography.* Paper presented at the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, CO.

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-Lib Magazine*, *8*(4). Retrieved January 22, 2010, from www.dlib.org/dlib/april02/weibel/04weibel.html

Eckersley, P., Egan, G., De Schutter, E., Yiyuan, T., Novak, M., Sebesta, V., et al. (2003). Neuroscience data and tool sharing. *Neuroinformatics*, *1*(2), 149–165.

Fedora Commons. (n.d.) *eScience/eResearch*. Retrieved February 2, 2010, from www.fedora-commons.org/about/examples/escienceeresearch

Foster, I., Vockler, J., Wilde, M., & Zhao, Y. (2002, June). *The virtual data grid: A new model and architecture for data-intensive collaboration*. Paper presented at the Workshop on Data Provenance and Derivation, Chicago, IL. Retrieved October 10, 2009, from people.cs.uchicago.edu/~yongzh/papers/CIDR.VDG.submitted.pdf

Fox, G. C., Guha, R., McMullen, D. F., Mustacoglu, A. F., Pierce, M. E., Topcu, A. E., et al. (2009). Web 2.0 for grids and e-science. In F. Davoli, N. Meyer, R. Pugliese, & S.

Zappatore (Eds.), *Grid enabled remote instrumentation* (pp. 409–431). New York: Springer.

Galloway, P. (2004). Preservation of digital objects. *Annual Review of Information Science and Technology*, *38*, 549–590.

Gardner, D., Goldberg, D., Grafstein, B., Robert, A., & Gardner, E. (2008). Terminology for neuroscience data discovery: Multi-tree syntax and investigator-derived semantics. *Neuroinformatics*, *6*(3), 161–174.

Gardner, D., Toga, A., Ascoli, G., Beatty, J., Brinkley, J., Dale, A., et al. (2003). Towards effective and rewarding data sharing. *Neuroinformatics*, *1*(3), 289–295.

Geambasu, R., Gribble, S. D., & Levy, H. M. (2009, June). *CloudViews: Communal data sharing in public clouds*. Paper presented at the HotCloud '09 Workshop, 2009 USENIX Annual Technical Conference, San Diego, CA. Retrieved June 12, 2009, from www.usenix.org/events/hotcloud09/tech/full_papers/geambasu.pdf

Geschwind, D. H. (2001). Sharing gene expression data: An array of options. *Nature Reviews Neuroscience*, *2*, 435–438.

Gladney, H. M. (2004). Trustworthy 100-year digital objects: Evidence after every witness is dead. *ACM Transactions on Information Systems*, *22*(3), 406–436.

Goslar, K., & Schill, A. (2004). Modeling contextual information using active data structures. In *Current Trends in Database Technology—EDBT 2004 Workshops* (pp. 325–334). Heidelberg, Germany: Springer.

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., & Heber, G. (2005). Scientific data management in the coming decade. *ACM SIGMOD Record*, *34*(4), 34–41.

Green, A. G., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories and domain specific data archives. *OCLC Systems & Services*, *23*(1), 35–53.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, *43*(4–5). Retrieved January 5, 2010, from tomgruber.org/writing/onto-design.htm

Hacker, T. J., & Wheeler, B. C. (2007). Making research cyberinfrastructure a strategic choice. *Educause Quarterly*, *2007*(1), 21–29.

Hammond, T., Moritz, T., & Agosti, D. (2008). The conservation knowledge commons: Putting biodiversity data and information to work for conservation. *Proceedings of the Twelfth Biennial Conference of the International Association for the Study of Commons*. Retrieved November 15, 2009, from hdl.handle.net/10535/2132

Hank, C., & Davidson, J. (2009). International data curation education action (IDEA) working group: A report from the second workshop of the IDEA. *D-Lib Magazine*, *15*(3/4). Retrieved January 24, 2010, from www.dlib.org/dlib/march09/hank/03hank.html

Harvard University Library. (2003). *Name resolution service: Introduction and use.* Cambridge, MA: The Library.

Hedstrom, M., & Montgomery, S. (1998). *Digital preservation needs and requirements in RLG member institutions*. Mountain View, CA: Research Libraries Group.

Heimbigner, D., & McLeod, D. (1985). A federated architecture for information management. *ACM Transactions on Information Systems*, *3*(3), 253–278.

Heinemann, A., Kangasharju, J., Lyardet, F., & Mühlhäuser, M. (2003). iClouds?: Peer-to-peer information sharing in mobile environments. In H. Kosch, L. Böszörményi, & H. Hellwagner (Eds.), *Proceedings of the 9th International Euro-Par Conference* (pp. 1038–1045). Berlin, Germany: Springer.

Helly, J. J., Elvins, T. T., Sutton, D., & Martinez, D. (1999). A method for interoperable digital libraries and data repositories. *Future Generation Computer Systems*, *16*(1), 21–28.

Helly, J. J., Elvins, T. T., Sutton, D., Martinez, D., Miller, S. E., Pickett, S., et al. (2002). Controlled publication of digital scientific data. *Communications of the ACM*, *45*(5), 97–101.

Henricksen, K., Indulska, J., & Rakotonirainy, A. (2002). Modeling context information in pervasive computing systems. In F. Mattern & M. Naghshineh (Eds.), *Proceedings of the First International Conference on Pervasive Computing* (pp. 167–180). Berlin, Germany: Springer.

Hess, C., & Ostrom, E. (2007). *Understanding knowledge as a commons*. Cambridge, MA: MIT Press.

Hey, T., & Trefethen, A. E. (2002). The UK e-Science core programme and the grid. *Future Generation Computer Systems*, *18*(8), 1017–1031.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, *1*(3), 134–140.

Hilgartner, S. (1995). Biomolecular databases: New communication regimes for biology? *Science Communication*, *17*(2), 240–263.

Hilgartner, S., & Brandt-Rauf, S. (1994). Data access, ownership, and control: Toward empirical studies of access practices. *Science Communication*, *15*(4), 355–372.

Interagency Working Group on Digital Data. (2009). *Harnessing the power of digital data for science and society*: Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council. Washington, DC: The Council.

Ives, Z. G., Halevy, A. Y., Mork, P., & Tatarinov, I. (2003). Piazza: Mediation and integration infrastructure for semantic web data. *Web Semantics*, *1*(2), 155–175.

Ives, Z., Khandelwal, N., Kapur, A., & Cakir, M. (2005, January). *Orchestra: Rapid, collaborative sharing of dynamic data*. Paper presented at the Second Biannual Conference on Innovative Data Systems Research, Asilomar, CA. Retrieved June 12, 2009, from www.cidrdb.org/cidr2005/papers/P09.pdf

Jaiswal, A. R., Giles, C. L., Mitra, P., & Wang, J. Z. (2006, November). *An architecture for creating collaborative semantically capable scientific data sharing infrastructures.* Paper presented at the 8th Annual ACM International Workshop on Web Information and Data Management, Arlington, VA.

Jankowski, N. W. (2007). Exploring e-science: An introduction. *Journal of Computer-Mediated Communication*, *12*(2, Article 10). Retrieved June 15, 2010, from jcmc.indiana.edu/vol12/issue2/jankowski.htm

Jin, J., & Ahn, G.-J. (2006a, June). *Role-based access management for ad-hoc collaborative sharing*. Paper presented at the Eleventh ACM Symposium on Access Control Models and Technologies, Lake Tahoe, CA. Retrieved June 6, 2009, from doi.acm.org/10.1145/1133058.1133086

Jin, J., & Ahn, G.-J. (2006b, November). *Towards secure information sharing and management in grid environments*. Paper presented at the 2006 International Conference on Collaborative Computing: Networking, Applications and Worksharing, Atlanta, GA. Retrieved April 29, 2009, from doi.ieeecomputersociety.org/10.1109/COLCOM.2006.361892

Karasti, H., Baker, K., & Halkola, E. (2005). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network. *Computer Supported Cooperative Work*, *15*(4), 321–358.

Karp, D., Carlin, S., Cook-Deegan, R., Ford, D., Geller, G., Glass, D., et al. (2008). Ethical and practical issues associated with aggregating databases. *PLoS Med*, *5*(9), e190.

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics: Re-shaping scientific practice. *Nature Reviews Genetics*, *10*, 331–335.

Kelton, K., Fleischmann, K. R., & Wallace, W. A. (2008). Trust in digital information. *Journal of the American Society for Information Science and Technology*, *59*(3), 363–374.

Kenney, A. R., McGovern, N. Y., Botticelli, P., Entlich, R., Lagoze, C., Payette, S., et al. (2002). Preservation risk management for web resources: Virtual remote control in Cornell's Project Prism. *D-Lib Magazine*, *8*(1). Retrieved January 22, 2010, from dlib.org/dlib/january02/kenney/01kenney.html

Kingsley, D. (2008, February). *Repositories, research and reporting: The conflict between institutional and disciplinary needs*. Paper presented at the VALA2008: Libraries, Technologies, and the Future Conference, Melbourne, Australia. Retrieved June 3, 2009, from www.valaconf.org.au/vala2008/papers2008/117_Kingsley_Final.pdf

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., et al. (2006). Data publication in the open access initiative. *Data Science Journal*, *5*, 79–83.

Koutrika, G. (2005). Heterogeneity in digital libraries: Two sides of the same coin. *DELOS Newsletter*, *3*. Retrieved April 3, 2009, from www.delos.info/index.php?option=com_content&task=view&id=411&Itemid=198

Kowalczyk, S. T. (2007). Digital preservation by design. In M. Raisinghani (Ed.), *Handbook of research on global information technology management in the digital economy* (pp. 405–431), New York: IGI Publishing.

Kunze, J. (2003, August). *Towards electronic persistence using ARK identifiers*. Paper presented at the Third European Conference on Digital Libraries Workshop on Web Archives, Trondheim, Norway. Retrieved October 6, 2009, from www.cdlib.org/inside/diglib/ark

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2005). Fedora: An architecture for complex objects and their relationships. *Journal of Digital Libraries*, *6*(2), 124–138. Retrieved October 5, 2009, from www.arxiv.org/abs/cs.DL/0501012

Lagoze, C., & Van De Sompel, H. (2007). Compound information objects: The OAI-ORE perspective. *Open Archives Initiative—Object Reuse and Exchange*. Retrieved October 5, 2009, from www.openarchives.org/ore/documents/CompoundObjects-200705.html

Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., Walters, W. H., & Kenney, A. R. (2000). *Risk management of digital information: A file format investigation*. Washington, DC: Council on Library and Information Resources. Retrieved September 24, 2009, from www.clir.org/PUBS/reports/pub93/pub93.pdf

Lesk, M. (2008). Recycling information: Science through data mining. *International Journal of Digital Curation*, *3*(1), 154–157.

Lubell, J., Rachuri, S., & Mani, M. (2008). Sustaining engineering informatics: Towards methods and metrics for digital curation. *International Journal of Digital Curation*, *3*(2), 59–73.

Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *Libraries and the Academy*, *3*(2), 327–336.

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities and relationships*. Bath, England: UKOLN.

Mann, R., Williams, R., Atkinson, M., Brodlie, K., Storkey, A., & Williams, C. (2002). Scientific data mining, integration, and visualization. *Technical Report UKeS-2002-06, National e-Science Centre*. Retrieved June 5, 2009, from citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.6966&rep=rep1&type=pdf

Maron, N. L., Smith, K. K., & Loy, M. (2009). *Sustaining digital resources: An on-the-ground view of projects today: Ithaka case studies in sustainability*. Bristol, England: JISC. Retrieved November 17, 2009, from www.ithaka.org/ithaka-s-r/strategy/ithaka-case-studies-in-sustainability/report/SCA_Ithaka_SustainingDigitalResources_Report.pdf

Massachusetts Institute of Technology. (2004). *DSpace system documentation: Architecture*. Retrieved September 26, 2009, from dlib.ionio.gr/software/dspace-1.2.1-2-docs/architecture.html

McGovern, N. (2007). A digital decade: Where have we been and where are we going in digital preservation? *RLG DigiNews*, *11*(1). Retrieved January 22, 2010, from hdl.handle.net/2027.42/60441

Michener, W. K. (2005). Meta-information concepts for ecological data management. *Ecological Informatics*, *1*(1), 3–7.

Michener, W. K., Beach, J., Bowers, S., Downey, L., Jones, M., Ludaescher, B., et al. (2005). Data integration and workflow solutions for ecology. In B. Ludaescher & L. Raschid (Eds.), *Data integration in the life sciences* (pp. 321– 324). Berlin, Germany: Springer.

Moore, R. W. (2008). Towards a theory of digital preservation. *International Journal of Digital Curation*, *1*(3), 63–75.

Moore, R. W., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., et al. (2000a). Collection-based persistent digital archives: Part 1. *D-Lib Magazine*, *6*(3). Retrieved January 24, 2010, from www.dlib.org/dlib/march00/moore/03moore-pt1.html

Moore, R. W., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., et al. (2000b). Collection-based persistent digital archives: Part 2. *D-Lib Magazine*, *6*(4). Retrieved January 24, 2010, from www.dlib.org/dlib/april00/moore/04moore-pt2.html

Moore, R. W., & Smith, M. (2007). Automated validation of trusted digital repository assessment criteria. *Journal of Digital Information*, *8*(2). Retrieved January 10, 2010, from journals.tdl.org/jodi/rt/printerFriendly/198/181

Morris, R., & Truskowski, B. (2003). The evolution of storage systems. *IBM Systems Journal*, *42*(2), 205–217.

Myers, J., Allison, T., Bittner, S., Didier, B., Frenklach, M., Green, W., et al. (2005). A collaborative informatics infrastructure for multi-scale science. *Cluster Computing*, *8*(4), 244–253.

National Aeronautics and Space Administration. (1986). *Report of the EOS Data Panel on the Data and Information System. NASA TM-87777, Earth Observing System*, Vol. IIa. Washington, DC: The Administration. Retrieved January 22, 2010, from ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19860021622_1986021622.pdf

National Institutes of Health. (2003). *NIH data sharing policy and implementation guidance*. Retrieved June 6, 2009, from grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

National Library of Medicine. (2009). The visible human project. Retrieved February 1, 2010, from www.nlm.nih.gov/research/visible/visible_human.html

Nelson, B. (2009). Data sharing: Empty archives. *Nature*, *461*, 160–163.

Niu, J., & Hedstrom, M. (2008, October). *Documentation of social science data.* Paper presented at the Annual Meeting of the American Society for Information Science and Technology, Columbus, OH.

Olson, G. M., Zimmerman, A., & Bos, N. (2008). *Scientific collaboration on the internet*. Cambridge, MA: MIT Press.

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *Journal of the American Society of Information Science and Technology*, *56*(11), 1140–1153.

Palmer, C. L., Cragin, M. H., & Hogan, T. P. (2004). Weak information work in scientific discovery. *Information Processing & Management*, *43*(3), 808–820.

Parr, C. S., & Cummings, M. P. (2005). Data sharing in ecology and evolution: Why not? *Trends in Ecology & Evolution*, *20*(7), 362–363.

Parsons, M. A., & Duerr, R. (2005). Designating user communities for scientific data: Challenges and solutions. *Data Science Journal*, *4*, 31–38.

Pearson, D. (2002, June). *The grid: Requirements for establishing the provenance of derived data*. Paper presented at the Workshop on Data Provenance and Derivation, Chicago, IL. Retrieved October 6, 2006, from people.cs.uchicago.edu/~yongzh/papers/Provenance_Requirements.doc

Pepe, A., Borgman, C. L., Wallis, J. C., & Mayernik, M. (2007, April). *Knitting a fabric of sensor data resources*. Paper presented at the International Conference on Information Processing in Sensor Networks. Cambridge, MA. Retrieved November 10, 2009, from polaris.gseis.ucla.edu/cborgman/pubs/pepe_ipsn_dsi_8.pdf

Piwowar, H. A., & Chapman, W. W. (2008). A review of journal policies for sharing research data. *Nature Precedings*. Retrieved January 22, 2010, from precedings.nature.com/documents/1700/version/1

Plale, B., Ramachandran, R., & Tanner, S. (2006, January/February). *Data management support for adaptive analysis and prediction of the atmosphere in LEAD.* Paper presented at the 22nd Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, Atlanta, GA.

Rajasekar, A. K., Marciano, R., & Moore, R. (1999, March). *Collection-based persistent archives*. Paper presented at the 16th IEEE Symposium on Mass Storage Systems, San Diego, CA. Retrieved January 22, 2010, from storageconference.org/STORAGE CONFERENCE/1999/papers/17rajase.pdf

Rajasekar, A. K., & Moore, R. W. (2008). Data and metadata collections for scientific applications. In B. Hertzberger, A. Hoekstra, & R. Williams (Eds.), *High-performance computing and networking* (pp. 72–80). Berlin: Springer.

Rajasekar, A. K., Wan, M., Moore, R. W., & Schroeder, W. (2004, June). *Data grid federation.* Paper presented at the International Conference on Parallel and Distributed Processing Techniques and Applications, Las Vegas, NV.

Rajasekar, A. K., Wan, M., Moore, R. W., & Schroeder, W. (2006). iRODS: Integrated rule-based data system. Retrieved January 22, 2010, from www.irods.org/pubs/DICE_irods-desc.pdf

Renear, A. H., Dolan, M., Trainor, K., & Cragin, M. (2009, October). *Towards a cross-disciplinary notion of data level in data curation*. Poster presented at the Annual Meeting of the American Society for Information Science and Technology, Vancouver, British Columbia, Canada. Retrieved January 22, 2010, from mail.asis.org/Conferences/AM09/posters/103.pdf

Research Collaboratory for Structural Bioinformatics Protein Data Bank. (2009). *An information portal to biological macromolecular structures*. Retrieved June 15, 2009, from www.pdb.org

Research Information Network. (2008). *To share or not to share: Publication and quality assurance of research data outputs. A report commissioned by the Research Information Network*. London: Research Information Network. Retrieved January 22, 2010, from www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs

Rew, R., & Davis, G. (1990). Data management: NetCDF: An interface for scientific data access. *IEEE Computer Graphics and Applications*, *10*(4), 76–82.

Ribes, D., & Finholt, T. A. (2007, November). *Tension across the scales: Planning infrastructure for the long term*. Paper presented at the International ACM Conference on Supporting Group Work, Sanibel Island, FL.

Ross, S., & McHugh, A. (2006). The role of evidence in establishing trust in repositories. *D-Lib Magazine*, *12*(7/8). Retrieved January 24, 2010, from www.dlib.org/dlib/july06/ross/07ross.html

Ruusalepp, R. (2008). *Comparative study of international approaches to enabling the sharing of research data*. London: JISC. Retrieved November 15, 2009, from www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf

Saltz, J. (2002, October). *Data provenance*. Paper presented at the Workshop on Data Provenance/Derivation Workshop, Chicago, IL. Retrieved October 10, 2006, from people.cs.uchicago.edu/%7Eyongzh/papers/ProvenanceJS10-02.doc

Scavo, T., & Welch, V. (2007, November). *A grid authorization model for science gateways*. Paper presented at the Annual International Workshop on Grid Computing Environments, Reno, NV. Retrieved January 30, 2010, from library.rit.edu/oajournals/index.php/gce/article/view/99

Schindler, U., & Diepenbroek, M. (2008). Generic XML-based framework for metadata portals. *Computers & Geosciences*, *34*(12), 1947–1955.

Schofield, P., Bubela, T., Weaver, T., Portilla, L., Brown, S., Hancock, J., et al. (2009). Post-publication sharing of data and tools. *Nature*, *461*(7261), 171–173.

Schuchardt, K., Pancerella, C., Rahn, L. A., Didier, B., Kodeboyina, D., Leahy, D., et al. (2007). Portal-based knowledge environment for collaborative science. *Concurrency and Computation: Practice and Experience*, *19*(12), 1703–1716.

Schwartz, C. (2000). Digital libraries: An overview. *Journal of Academic Librarianship*, *26*(6), 385–393.

Semantic Counteroperability Community of Practice. (2008). Establish federated governance. Retrieved January 22, 2010, from semanticommunity.wik.is/Best_Practices/Enterprise_Mashup:_A_Practical_Guide_to_Federal_Service_Oriented_Architecture/Section_4:_Keys_to_Federal_SOA_Implementation/4.1_Keys_to_Implementing_the_Service-Oriented_Enterprise/4.1.7_Establish_Federated_Governance

Shafer, K. E., Weibel, S. L., & Jul, E. (2001). The PURL project. *Journal of Library Administration*, *34*(1), 123–125.

Shankar, K. (2007). Order from chaos: The poetics and pragmatics of scientific recordkeeping. *Journal of the American Society for Information Science and Technology*, *58*(10), 1457–1466.

Shiffrin, R. M., & Börner, K. (2004). Mapping knowledge domains. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5183–5185.

Simmhan, Y., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record*, *34*(3), 31–36.

Simmhan, Y., Plale, B., & Gannon, D. (2006, September). *A performance evaluation of the Karma provenance framework for scientific workflows.* Paper presented at the International Conference on Web Service, Chicago, IL.

Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., et al. (2003). DSpace: An open source dynamic digital repository. *D-Lib Magazine*, *9*(1). Retrieved January 22, 2010, from www.dlib.org/dlib/january03/smith/01smith.html

Sonnenwald, D. H., Whitton, M. C., & Maglaughlin, K. (2008). Evaluation of a scientific collaboratory system: Investigating a collaboratory's potential before deployment. In G. Olson, A. Zimmerman, & N. Bos. (Eds.), *Scientific collaboration on the internet* (pp. 171–194). Cambridge, MA: MIT Press.

Stanescu, A. (2005). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *OCLC Systems & Services*, *21*(1), 61–81.

Steinhart, G. (2007). DataStaR: An institutional approach to research data curation. *IASSIST Quarterly, 31*(3–4), 34–39. Retrieved June 5, 2009, from hdl.handle.net/1813/12668

Toronto International Data Release Workshop. (2009). Prepublication data sharing. *Nature*, *461*(10), 168.

Treloar, A., Groenewegen, D., & Harboe-Ree, C. (2007). The data curation continuum: Managing data objects in institutional repositories. *D-Lib Magazine*, *13*(9). Retrieved June 6, 2009, from www.dlib.org/dlib/september07/treloar/09treloar.html

Treloar, A., & Wilkinson, R. (2008, December). *Rethinking metadata creation and management in a data-driven research world.* Paper presented at the Fourth International Conference on eScience, Indianapolis, IN.

U.S. National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century: Report of the National Science Board*. Arlington, VA: National Science Foundation.

U.S. Social Security Administration (2009). *Government information exchange glossary*. Retrieved June 5, 2009, from www.ssa.gov/gix/definitions.html

Van House, N. (2002). Digital libraries and practices of trust: Networked biodiversity information. *Social Epistemology*, *16*(1), 99–114.

Van House, N., Butler, M., & Schiff, L. (1998). Cooperative knowledge work and practices of trust: Sharing environmental planning datasets. *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*, 335–343.

Vannier, M. W., Staab, E. V., & Clarke, L. P. (2003). Matching clinical and biological needs with emerging imaging technologies. *Proceedings of the IEEE*, *91*(10), 1562–1573.

Venugopal, S., Buyya, R., & Ramamohanarao, K. (2006). A taxonomy of data grids for distributed data sharing, management, and processing. *ACM Computing Surveys*, *38*(1). Retrieved April 21, 2009, from portal.acm.org/citation.cfm?id=1132952.1132955#

Vogeli, C., Yucel, R., Bendavid, E., Jones, L., Anderson, M., Louis, K. S., et al. (2006). Data withholding and the next generation of scientists: Results of a national survey. *Academy of Medicine*, *81*(2), 128–136.

Wallis, J., Borgman, C., Mayernik, M., & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. *International Journal of Digital Curation*, *1*(3), 114–126.

Waters, D., & Garrett, J. (1996). *Preserving digital information: Report of the task force on archiving of digital information*. Washington, DC: Commission on Preservation and Access and Research Libraries Group.

Weill, P., & Ross, J. W. (2004). *IT governance: How top performers manage IT decision rights for superior results*. Boston: Harvard Business School Press.

Westbrook, J., Feng, Z., Chen, L., Yang, H., & Berman, H. M. (2003). The Protein Data Bank and structural genomics. *Nucleic Acids Research*, *31*(1), 489–491.

Westbrook, J., Ito, N., Nakamura, H., Henrick, K., & Berman, H. M. (2005). PDBML: The representation of archival macromolecular structure data in XML. *Structural Bioinformatics*, *21*(7), 988–992.

Whyte, A., Job, D., Giles, S., & Lawrie, S. (2008). Meeting curation challenges in a neuroimaging group. *International Journal of Digital Curation*, *3*(1), 171–181.

Willinsky, J. (2006). *The access principle: The case for open access to research and scholarship*. Cambridge, MA: MIT Press.

World Wide Web Consortium. (2001). *URIs, URLs, and URNs: Clarifications and recommendations 1.0. Report from the Joint W3C/IETF URI Planning Interest Group, W3C Note 21.* Retrieved October 6, 2006, from www.w3.org/TR/uri-clarification

Wright, D. J. (2009). Spatial data infrastructures for coastal environments. In X. Yang (Ed.), *Remote sensing and geospatial technologies for coastal ecosystem assessment and management* (pp. 91–112). Heidelberg, Germany: Springer.

Wroe, C., Goble, C., Greenwood, M., Lord, P., Miles, S., Papay, J., et al. (2004). Automating experiments using semantic data in a bioinformatics grid. *Intelligent Systems*, *19*(1), 48–55.

Zhang, J., Altintas, I., Tao, J., Liu, X., Pennington, D. D., & Michener, W. K. (2006). Integrating data grid and web services for e-science applications: A case study of exploring species distributions. *Second IEEE International Conference on e-Science and Grid Computing*, 31–39.

Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data: Experiences of ecologists*. Unpublished doctoral dissertation, University of Michigan.

Zuccala, A. (2009). The layperson and open access. *Annual Review of Information Science and Technology*, *43*, 359–396.