e-Science Data Environments:
A View from the Lab Floor

Stacy T. Kowalczyk

Submitted to the faculty of the University Graduate School in partial fulfillment
of the requirements for the degree Doctor of Philosophy in the
School of Library and Information Science,
Indiana University
December 2011

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Katy Börner, Ph.D., Chair

_____

Debora Shaw, Ph.D.

_____

Alan Dennis, Ph.D.

_____

Staša Milojević, Ph.D.

_____

Beth Plale, Ph.D.

December 12, 2011

Acknowledgements

I am deeply grateful for the support of my faculty advisors, my fellow students, and my family.

Dr. Katy Börner, my advisor, for her unfailing encouragement.

Dr. Ralf Shaw, for her most excellent comments and suggestions.

Dr. Allan Dennis, for a wonderful research opportunity early in my academic career.

Dr. Staša Milojević, for her advice to tell my story.

Dr. Beth Plale, for the opportunity to be part of a research environment that is both nurturing and challenging.

Shannon Oltmann, my friend and writing partner, for her willingness to read and re-read my dissertation, offer advice on content and style, and provide a bi-weekly deadline that made me write.

George and Janice Travers, my parents, for a lifetime of love and encouragement.

Kerry, Megan, Zach, and Jake; Will, Brendan, Nora, and Maggie, my family, for their unbridled enthusiasm for my studies.

Andrew Kowalczyk, my dear husband and partner, for living with a crazy woman for the past 5 years, for never minding the mess, for giving me quiet time on weekend mornings, for loving me.

Stacy T. Kowalczyk

e-Science Data Environments: A View from the Lab Floor

Preserving scientific digital data and ensuring its continued access has emerged as a major initiative for both funding agencies and academic institutions. Digital preservation, the study of the processes, organizations, and technologies needed to maintain scientific digital data over time, is a multidisciplinary field that draws on the literature from library and information science, informatics/computer science, and domain sciences such as biology, geology, and environmental sciences. This dissertation develops and tests a new theoretical model for the preservation of scientific data concerning the research practices of scientists, adds to knowledge about the lifecycle of research data, and the related antecedents, barriers, and threats to data preservation. This research is based on a mixed-method approach. An initial study was conducted using case study analytical techniques at the individual level. Insights from these case studies were combined with grounded theory in order to develop a novel model of the e-Science Data Environment. A broad-based quantitative survey was then constructed to test and extend the components of the model. The major contributions of these research initiatives are the creation of the e-Science Data Environment, a data lifecycle that provides a generalized model of the research process and a theoretical basis to better explain and predict both the antecedents and barriers to preservation.

Table of Contents

List of Tables

List of Figures

e-Science Data Environments: A View from the Lab Floor

1 Introduction

Computer-based technology has fundamentally changed both the practice of science and the output of science (Gray, Szalay, Thakar, Stoughton, & vandenBerg, 2002; Housewright & Schonfeld, 2008). The practice of science is becoming increasingly interdependent, interdisciplinary, and data driven (Anderson, 2004; Hey & Trefethen, 2003). Scientific output, in the form of digital data, is being created at an increasingly fast pace. The dramatic growth of scientific digital data generation is referred to as the "data deluge" in both the scientific literature (Borgman, Wallis, & Enyedy, 2007; Gershon, 2002: Hey & Trefethen, 2003; Jirotka, Procter, Rodden, & Bowker, 2006) and popular technology venues such as magazines (Anderson, 2008) and blogs (Losh, 2010; Reed, 2010).

Preserving scientific digital data and ensuring its continued access, has emerged as a major priority for funding agencies (Association of Research Libraries [ARL], 2006; Atkins, 2003; Hedstrom, Dawes, Fleischhauer, Gray, Lynch et al., 2003; Interagency Working Group on Digital Data [IWGDD], 2009; Lord & Macdonald, 2003; Lyon, 2007; National Science Board [NSB], 2005) and academic institutions (Davis & Connolly, 2007). Over the past ten years, numerous workshops have been held, position papers have been written, and reports have been published describing the value of preserving digital scientific digital data. Multiple reasons justify this growing interest in digital data preservation: the data itself has significant scientific value; it can be reused to fuel new ideas and insights (ARL, 2006; Atkins, 2003; IWGDD, 2009;

NSB, 2005); it is an integral part of the scientific record as evidence of the rhetorical structure of scholarly communication (Rusbridge, 2007; Swan & Brown, 2008); and it is necessary for replication and validation of scientific results (Swan & Brown, 2008).

Scientific digital data has significant economic value. Because the research investment that is typically necessary to create the data is costly, it is not a commodity that should be consumed in a single use; rather, data is intellectual capital, an important and invaluable resource that can be used repeatedly (ARL, 2006; Atkins, 2003; IWGDD, 2009; Lord & Macdonald, 2003;Rumsey, 2010). Funding agencies promote shared access to data as the wisest use of public resources, which can mitigate repetitive collection of expensive or sensitive data by making experimental and observational results available to the scientific community (Atkins, 2003; Lord & Macdonald, 2003). Scientific digital data is a generalized good; society benefits both directly and indirectly when this data is available for citizen scientists, for teaching, for commercial reuse, and for policy development (IWGDD, 2009; Lord & Macdonald, 2003; Rumsey, 2010).

Providing long-term access to scientific digital data has a number of challenges. Digital data requires constant and perpetual maintenance (Hedstrom &Montgomery, 1998). Technologies change; equipment ages; software is superseded. Digital data is not fixed and can easily be changed, either intentionally or unintentionally (Gladney, 2004). Securing data is an essential to ensuring its quality and value.

Long-term access is difficult to define because preservation time frames vary widely. Long-term may mean the length of a single project (Beagrie, 2006). Or, as a report from the National Science Foundation and the Library of Congress describes, long-term can mean several decades, generations, or centuries (Hedstrom et al., 2003). Long-term can even imply unlimited

time periods (IWGDD, 2009; Lord & Macdonald, 2003). The Atkins (2003) report states, "absent systemic archiving and curation of intermediate research results… data gathered at great expense will be lost" (p. 11).

1.1 Defining Digital Preservation

Digital preservation, the processes and technologies needed to maintain digital materials over time, is a multidisciplinary field that draws strongly on the fields of library and information science (LIS) and informatics/computer science (CS). Each discipline brings different perspectives to the field, which can be reflected in the differing definitions of the field itself. Baker, Keeton, and Martin (2005) use a CS perspective in their definition of digital preservation – storing immutable data over long periods of time. The CS perspective is clearly focused on the technology.

In LIS, digital preservation is defined as "the managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents" (Research Libraries Group & OCLC, 2002, p. 11). Another LIS definition of digital preservation is "the planning, resource allocation, and application of preservation methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable" (Hedstrom, 1997, p. 190). The LIS definitions of digital preservation highlight the requirement of management, stewardship, and long-term availability and usability of the data. In library and information science, the definition of preservation has been recently expanding to include digital archiving and digital curation. These definitions are still evolving and are regularly used interchangeably, much to the detriment of clear and concise communication (Beagrie, 2006).

In order to address the growing need for specificity and clarity related to the issues

important to digital preservation, Lord, Macdonald, Lyon, and Giaretta (2004) developed a set of differentiated definitions:

> **Curation:** The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and other published materials.
>
> **Archiving:** A curation activity, which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.
>
> **Preservation**: An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology (Lord et al., 2004, p. 1).

The granularity of these definitions indicates a maturity of some of the thinking about digital preservation. However, these definitions are not widely used. Identifying "preservation" as a sub-activity of archiving, which is itself a sub-activity of curation, constrains the most commonly used general word for the total set of activities. This dissertation will use the term "preservation" for the total set of activities described above but will additional attempt to disambiguate between the definitions set forth by Lord and colleagues (2004) when the literature is vague.

1.2 Structure of the Dissertation

This dissertation develops and tests a new theoretical model for the preservation of scientific data. This work is motivated and set within the research literature discussed in Chapter 2. Chapter 3 presents the results of a formative analysis for this work and lays the foundation for the design of a study that tests and enhances this theory. Chapter 4 describes a survey designed to both to refine and extend the generalizability of the e-Science Data Environment model. Chapter 5 presents the results of the survey. Chapter 6 describes the new theoretical model of the antecedents and barriers to preservation in the e-Science Data Environment. Chapter 7 summarizes the significance of the research.

2 Literature Review

2.1 The Nature of Data

Preserving scientific data begins with data. Data "refer[s] to 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" (NSB, 2005, p. 9). Data is an inherently collective word as it comes in sets – the collation of many individual datum. In most scientific research, researchers create or use a number of data sets or databases. These sets of data sets are referred to as data collections (Kowalczyk & Shankar, 2011). Data collections are defined as the infrastructure, organizations, and individuals needed to provide persistent access to stored data (NSB, 2005).

Data collections have been described by a three-layer typology that is organized by size and scope: research data collections, community data collections, and reference data collections (NSB, 2005). Research data collections refer to the output of a single researcher or lab during the course of a specific research project. This collection may use the data standards of its community and may have use beyond its own original purpose. Community data collections generally serve a well defined area of research. Often, standards are developed by the community to support the collection. At the highest level, reference data collections are broad in scope, widely disseminated, and well funded collections that support the research needs of many communities (NSB, 2005). There is little evidence that researchers individually or as communities identify themselves in this taxonomy or use these terms to describe their own activities and repositories (Kowalczyk & Shankar, 2011). Nevertheless, the framework is used

widely in the digital preservation literature (Beagrie, Beagrie, & Rowlands, 2009; Green & Gutmann, 2007; Karasti & Baker, 2008) and provides a useful way of describing the functions, structure, and organizational dimensions of data collections and the resulting changes that arise from scaling up and expanding the scope of use.

Data can be dynamic and subject to change; or data can be stable and non-changing (Lord & Macdonald, 2003; Lyon, 2007). The more volatile the data, the more difficult to manage, store, and access (Guy, Kunszt, Laure, Stockinger, & Stockinger, 2002). But even stable data changes as it is processed. Data moves through stages from its original raw state via processing and analysis to a final state (Helliwell, Strickland, & McMahon, 2006; Lyon, 2007; National Aeronautics and Space Administration [NASA], 1986; Witt, Carlson, Brandt, & Cragin, 2009). Liu and Chi (2002) argue that data evolves through four stages: collection, organization, presentation, and application. This evolution can be described as a lifecycle.

2.2 Lifecycles and Data Preservation

Lifecycle models can be used to represent the flow, relationships, and transitions of major components of large systems (Humphrey, 2006). Lifecycles provide an important and useful framework for understanding data preservation because active intervention early in that lifecycle is considered essential for success (Beagrie, 2006; Rice, 2007; Rumsey, 2010). In their work on lifecycles, Wallis, Borgman, Mayernik, and Pepe (2008) argue that the number of individuals and institutions involved at each stage of the lifecycle increases as the complexity of the data increases. Data lifecycles are path dependent (Rumsey, 2010). The cumulative weight of decisions made at each stage determines what is available at the next, how it is handled, and the purposes for which it is useful (Wallis et al., 2008). Iwata (2008) posits that the time constants

of data lifecycles are becoming shorter and that the diversity of stakeholders and complexities of data are increasing. Although the data itself may be more or less easily depicted through various descriptive processes, documenting the decisions at each stage of the lifecycle is more problematic and less easily automated (Borgman, 2007; Higgins, 2008; Wallis et al., 2008).

Lifecycles are used in a number of fields, such as quality research (Fendt, 2004; Dasu, Vesonder, &Wright, 2003; Levitin & Redman, 1993; Otto, Wende, Schmidt, & Osl, 2007; Wang, Storey, & Firth, 1995), data warehouse research (Inmon, Strauss, & Neushloss, 2008: Mathieu, & Khalil, 1998), knowledge transfer research (Humphrey, 2006), and data management (Loshin, 2009). Some technologies have inspired specific data life cycles including Radio Frequency Identification (RFID) (Niederman, Mathieu, Morley, & Kwon, 2007) and ambient intelligence environments (e.g. smart phones) (Anciaux, Van Heerde, Feng, & Apers, 2006).

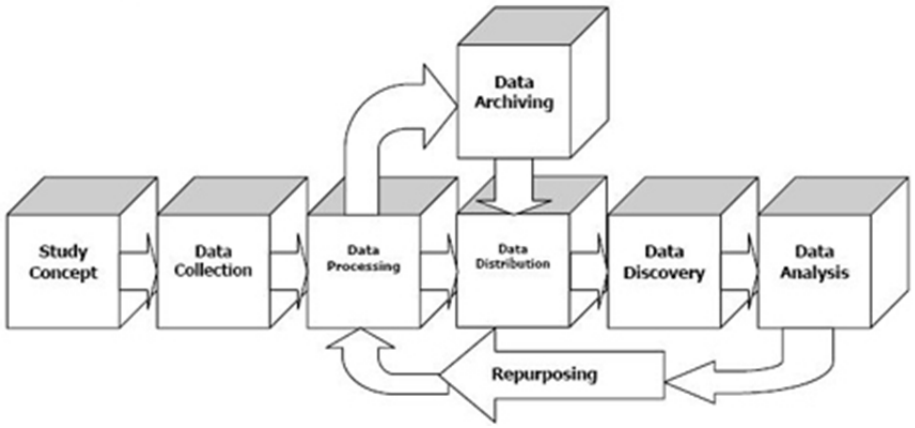


Figure 1. DDI Lifecycle

(Thomas, Gregory, & Piazza, 2005)

Within the digital preservation literature, lifecycles are either generic in that they pertain to an entire domain; or they are specific because they pertain to a particular lab or project. The social science community has developed two generic lifecycles to describe the basic research model. The Data Documentation Initiative (DDI), a metadata standard for data description used

in the social sciences, is based on a lifecycle model (Green, 2008; Martinez, 2008). As seen in Figure 1, the model describes an eight-stage linear research model with cyclical reuse and archiving; it is based on the concept of production, that is, the actual creation of the data (Ryssevik, 2001; Vardigan, Heus, & Thomas, 2008). The conception of the study and data collection phases are considered pre-production stages; the data processing is considered production; data archiving and data distribution are considered post-production activities; and data discovery, analysis, and repurposing are considered secondary use (Green, 2008). The second generic lifecycle model (see Figure 2) is a five-stage cycle that includes discovery and planning, initial data collection, final data preparation and analysis, publication and sharing, and long-term management (Green & Gutmann, 2007).

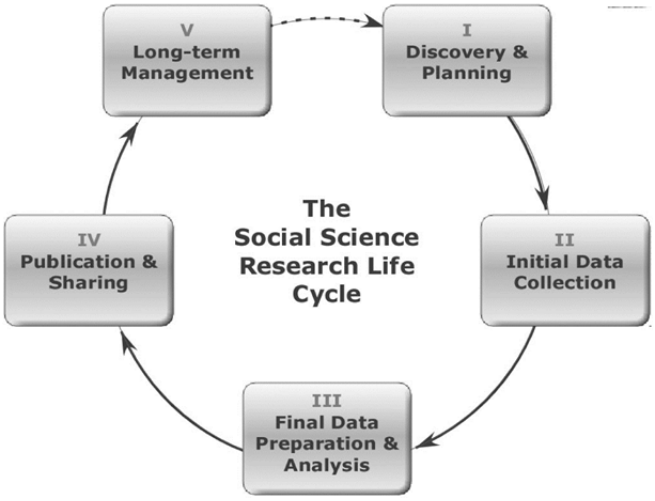


Figure 2. Research Life Cycle

(Green & Gutmann, 2007)

The Library of Congress has developed a generic preservation lifecycle that compares the traditional preservation model of non-digital materials to an emerging model for digital preservation (see Figure 3). The traditional model focuses on "fixing information to physical objects; the conservation of the physical objects becomes the mode of preserving the

information" (Rumsey, 2010, p. 29). In the digital preservation lifecycle, preservation actions are initiated at every stage of the lifecycle. Preservation actions are of particular import when responsibility for the data changes hands. These handoffs, from the creator to the curator, involve technology and policy that can affect the long-term viability of the data to be preserved (Rumsey, 2010).

Figure 3. Traditional and Digital Preservation Lifecycles

(Library of Congress, 2011)

Domain or project-specific lifecycles have been developed from case studies (Borgman, 2007; Higgins, 2008; Long, Mantey, Wittenbrink, Haining, & Montague, 1995; Wallis et al., 2008). Lifecycles differ widely between different scientific domains (Borgman, Bowker, Finholt, & Wallis, 2009). One example of a domain-specific lifecycle is the nine-stage data lifecycle of the Center for Embedded Networked Sensing (CENS) (Borgman, 2007; Higgins, 2008; Wallis et al., 2008). Although depicted as a cycle (see Figure 4), it portrays a primarily linear process; the experimental design leads to data capture, which leads to a process of data cleaning, integration and derivation, which leads to analysis, and so on. The CENS lifecycle was

developed as part of a larger research project to design an architecture for a large archive and delivery system for embedded sensing data.

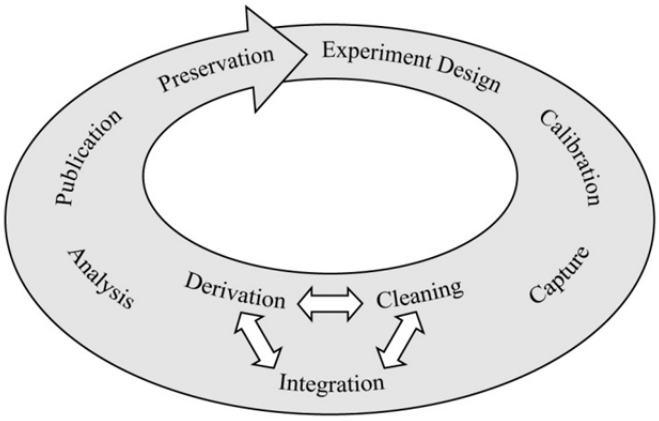


Figure 4. CENS Lifecycle

(Wallis et al., 2008)

Long-term data management needs to be part of the lifecycle of data (Green & Gutmann, 2007); but it is only a minor part of most data and research lifecycles (Pryor & Donnelly, 2009). The lifecycles just examined do indeed include data management using alternative terms such as data preservation or data archiving. However, data management, data preservation, or data archiving are final steps, almost an afterthought. There is no indication of any data preparation for the preservation or archiving activities. The preservation lifecycle does not significantly improve on the scientific lifecycles. Although preservation actions are indicated, there is no indication of what those actions could or should be.

In these lifecycles, there is no discussion about intermediate file management. There is no discussion of assessment, the process of determining which files to preserve. There is little discussion about what is meant by preservation/archiving. These life cycles leave a number of questions unanswered: What actually is preserved – all files, some files, or final files? Where will this data be preserved/archived – on lab storage devices, in institutional repositories, or in

community supported research collections? Who will be the long-term caretakers or curators of this data – the individual researcher, the institution at which the research was conducted, or the funding organization? These open questions are a symptom of a larger problem: the paucity of theory to explain and model the preservation of scientific data.

2.3 Preservation Risk Factors

Assessing, describing, and managing risk is a significant component of preservation infrastructure management (Kenney, McGovern, Botticelli, Entlich, Lagoze, & Payette, 2002; Stanescu, 2005; Whyte, Job, Giles, & Lawrie, 2008). The risks associated with preservation infrastructures have been defined through the literature as threats to preservation (Altman, Adams, Crabtree, Donakowski, Maynard, Pienta, et al., 2009; Baker, Keeton, & Martin, 2005; Barateiro, Antunes, Cabral, Borbinha, & Rodrigues, 2008; Hunter & Choudhury, 2004; Rosenthal, Robertson, Lipkis, Reich, & Morabito, 2005; Rosenthal, Roussopoulos, Giuli, Maniatis, & Baker, 2004). In the literature, these threats have been categorized in a number of ways: component failures, management failures, disasters, and attacks (Barateiro et al., 2008); physical threats, technology threats, human threats, and institutional threats (Altman et al., 2009); economic threat, human error, disasters and attacks (Rosenthal et al., 2004). But at the most elemental level, threats to reliable preservation are either technology-based or human based. Table 1, below, provides a summary of relevant issues raised in the literature about various technology-based threats to digital preservation. This literature is organized according to thematic emphasis of the authors.

Table 1. Technology-based Threats to Preservation

Threat	Authors
Hardware failures	Altman et al., 2009; Baker et al., 2005; Barateiro et al., 2008; Rosenthal et al., 2005
Media flaws	Barateiro et al., 2008
Massive storage failure	Altman et al., 2009; Baker et al., 2005; Rosenthal et al., 2005
Intermittent failures ("bit rot")	Baker et al., 2005
Network services failures	Barateiro et al., 2008; Rosenthal et al., 2005
Software failures	Altman et al., 2009; Baker et al., 2005; Barateiro et al., 2008; Rosenthal et al., 2005

Technology-based threats are defined as failures where the hardware, software, or storage media did not perform as expected. These failures can include complete stoppage, intermittent stoppage, data corruption introduced by hardware or software errors/failures, loss of services, or manufacturing flaws in the storage media. Data corruption is the most significant threat to preservation. Preserving "bad" data is hardly preservation. Thus, fixity is a significant data preservation issue (Gladney, 2004; Hedstrom & Montgomery, 1998). Some storage devices have internal fixity functions (Rabinovici-Cohen, Factor, Naor, Ramati, Reshef, Ronen, et al., 2008), but it is often a software responsibility. In preservation, that software is generally the repository. 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 file. Checksums or signatures created before the files are placed into the repository are better able to insure integrity and fixity if they are stored in the repository and validated by a process that calculates the checksum for each file and compares it

to the saved checksum on a regular schedule, reporting errors to the repository managers (Kowalczyk, 2008).

Many of the threats to preservation have been previously categorized as technology faults when, in practice, these threats are really human failures – failures of attention, failures of management, and failures of planning. Table 2, below, provides a summary of relevant issues raised in the literature about various human-based threats to digital preservation. This literature is organized according to thematic emphasis of the authors. (see Table 2). Data management, as discussed in section 2.5.1, is a set of tasks and activities to plan for system backups, contingencies, process resumption, hardware and network redundancy, automatic failover, and site mirroring – all efforts to avert problems. Some problems may come from non-human sources, such as bad weather precipitating power failures or flooding; however, the primary data management failures are human. Data deletion, whether accidental or purposeful, and operator errors, such as overwriting tapes, misnaming files, and linking incorrect metadata, are the all-too-frequent human failures that threaten preservation. Threats categorized as institutional are also human based. Changing priorities and the attending financial choices as well as lack of management attention have a higher probability of threatening an institution's commitment to preservation than do technology threats.

Obsolescence of hardware, software, and media has been categorized as a technology risk (Altman et al., 2009; Barateiro et al., 2008; Rosenthal et al., 2005). Since obsolescence in inherent to technology, it cannot be considered some vague future threat but a surety. Planning for obsolescence is a specific management task. It is the lack of management that is the threat. Because management is a human activity, obsolescence must be considered a human risk.

Table 2. Human-based Threats to Preservation

Threat	Authors
Erasure error	Baker et al., 2005
Loss of context	Baker et al., 2005
Operator error	Altman et al., 2009; Barateiro et al., 2008; Rosenthal et al., 2005; Rosenthal et al., 2004
Lack of Disaster preparedness	Barateiro et al., 2008; Rosenthal et al., 2005; Rosenthal et al., 2004
Communication errors	Barateiro et al., 2008; Rosenthal et al., 2005
Media obsolescence	Baker et al., 2005; Barateiro et al., 2008; Hunter & Choudhury, 2004; Rosenthal et al., 2005
Format obsolescence	Altman et al., 2009; Baker et al., 2005; Barateiro et al., 2008
Software obsolescence	Baker et al., 2005; Barateiro et al., 2008; Hunter & Choudhury, 2004; Rosenthal et al., 2005
Hardware obsolescence	Hunter & Choudhury, 2004; Rosenthal et al., 2005
Funding / Economic Failure	Altman et al., 2009; Baker et al., 2005; Barateiro et al., 2008; Hunter & Choudhury, 2004; Rosenthal et al., 2005; Rosenthal et al., 2004
Institutional failure	Baker et al., 2005; Barateiro et al., 2008; Hunter & Choudhury, 2004; Rosenthal et al., 2005
Mission change	Altman et al., 2009
Legal regime change	Altman et al., 2009
Malicious attack – either internal or external	Altman et al., 2009; Baker et al., 2005; Barateiro et al., 2008; Rosenthal et al., 2005; Rosenthal et al., 2004

Baker and colleagues (2005) discovered that two generally accepted data management truisms that greatly affect the perception of preservation risks are actually fallacies. The first assumption is that all faults can be easily and quickly detected and repaired. They found that faults can be visible, detected, and fixed immediately, or latent, undetected, and dormant. Latent faults result in data that is either irretrievable or corrupt. The second assumption is that faults happen independently. They found that failures are not independent but happen in multiple, cascading, and compounding occurrences (Baker et al., 2005).

Research on lost data is sparse. Holzner, Igo-Kemenes, and Mele (2009) found that approximately 40% of the high-energy physicists surveyed think that they have lost important data in the past. But few of these losses are documented. Two of the most famous, or perhaps infamous, examples of lost data are both from NASA. The first of these is data collected during the Apollo 11, 12, and 14 moon missions from instrumentation for collecting lunar surface environmental information. The 173 tapes of that data were misplaced by the University of Sydney's data center before they were archived or documented by NASA. In 2008, researchers realized that the data on lunar atmospheric dust could be useful for future lunar exploration research. The tapes were located but the tape drives needed to read the data were no longer available (MacBean, 2008). The second example is a similar story. The tapes of Neil Armstrong's first walk on the moon, along with approximately 700 other data tapes of Apollo data, were withdrawn from the National Archives in 1984 by NASA's Goddard Space Center in Maryland. These tapes are now missing (Macey, 2006). Significant data, both historical and scientific, was lost. These two losses of valuable data can both be attributed to human error. As Baker and colleagues (2005) described, multiple, cascading, and compounding errors occurred: data was not managed; technology was not monitored; policies were ignored.

Preservation threats have been analyzed and discussed within the context of a preservation infrastructure, an archival system. As such, they are considered dangers to preserved data. But for scientists who have not yet preserved their data, these threats become barriers, obstacles to be overcome.

2.4 Emerging Theories of Digital Preservation

Ross (2007) posits that, despite more than twenty years of research, the field of digital curation and preservation has developed few "actual theories, methods and technologies that can either foster or ensure digital longevity" (p.1). Most of the research in digital preservation is systems-centered, focusing on "component technologies and integration to realize information environments that are dynamic and flexible" (National Science Foundation, 1998). One of the few emergent theories in digital preservations takes a systems-centered approach by focusing primarily on the technical issues of digital preservation repositories (Moore, 2008).

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 provenance so future systems, as of yet unimagined, will be able to interpret and display the information. The conversation with the past provides characterizations of prior preservation processes and preservation management policies. Motivated by the need to create data and information management technologies as persistent archives, this theory defines a preservation environment that is a system, a middleware layer, which protects data from technological change and obsolescence by providing standard preservation operations on persistent objects via a set of machine actionable policies (Watry, 2007; Moore, 2008; Moore & Smith, 2007). This nascent theory presumes that the data has been either created in a

Cyberinfrastructure environment or pushed into a preservation environment; it does not address the antecedents to preservation. These antecedents, however, are crucial to the act of preservation and are the focus of this dissertation.

2.5 Antecedents to Preservation

The most salient antecedents to preservation, the actions that must precede preservation, are obvious. Data must be created. It must be knowable; that is, it must be able to be found and accessed. Thus, the data needs to be described in such a way that it can be understood within its context. It must be available when preservation is deemed necessary. There must be a technical environment in which the data will be preserved.

Data must be managed from its creation (Lynch, 2008). Institutions, data centers, users, funders, data creators, and publishers all have roles, rights, and responsibilities for curating, archiving, and preserving data (Hey & Trefethen, 2003; Lyon, 2007). But currently, all of the antecedents to preservation are the responsibility of the scientists who created the data. The scientist is responsible to manage data for life of the project, to meet standards of good practice, and to "work up data" for use by others (Lyon, 2007, p. 9). The burden is on the scientist to manage the data, create the contextual metadata, and determine a final disposition of the data.

2.5.1 Data Management

Data management is the term used to describe the collective tasks to insure the long term archiving of and continuing access to data, including backups, contingency planning, process resumption planning, hardware and network redundancy, automatic failover, and site mirroring. Rusbridge (2007) claims that data management is a discipline that requires the necessary context information and associated documentation needed to ensure successful use and re-use of data. It

is a dynamic process that needs to be mindful of the entire data lifecycle (Hank & Davidson, 2009; Rice, 2007). As the amount of data increases, so does the complexity and resources required to manage the data (Pritchard, Anand, & Carver, 2005).

Established and well funded data collections often have dedicated data management staff (Karasti, Baker, & Halkola, 2006) and large data centers to manage the petabyte datasets (Gray, Liu, Nieto-Santisteban, Szalay, DeWitt, & Heber, 2005; Hey & Trefethen, 2003). But in the main, it is individual scientists' responsibility to manage their data (Henty, Weaver, Bradbury, & Porter, 2008; Lynch, 2008; Lyon, 2007; Pritchard, Anand, & Carver, 2005). Scientists expect to manage their data and understand the importance of good data management but often are unsure of how best to implement good data management practices (Henty et al., 2008; Pritchard, Anand, & Carver, 2005; Pryor & Donnelly, 2009). Without clear direction from funding agencies, researchers are left to create their own guidelines (Jones, Ball, & Ekmekcioglu, 2008; Marcus, Ball, Delserone, Hribar, & Loftus, 2007). As discussed in section 2.2, decisions made at one stage of the research lifecycle affect the range of options available at a later stage (Rumsey, 2010). Thus, decisions made as data is created are of the utmost importance because they influence all subsequent decisions.

2.5.2 Contextual Metadata

Metadata is a key factor for data preservation (Anderson, 2004: Hey & Trefethen, 2003; National Information Standards Organization [NISO], 2008; Rajasekar & Moore, 2001; Swan & Brown, 2008), for replicating results in the peer review process (Vardigan, Heus, & Thomas, 2008), for data reuse (Gray et al, 2005; Hey & Trefethen, 2003; Lyon, 2007), and for creating knowledge from data (Hey & Trefethen, 2003). Well managed data without metadata could be useless (Rumsey, 2010). Lesk (2008) contends that preservation, the long-term persistence of

data, is tightly coupled with access; funding for preservation and curation activities will be based on the perceived usefulness and accessibility to the data. Access requires metadata. And metadata is expensive; there is usually a direct relationship between the cost of metadata creation and the benefit to the user (NISO, 2008). Creating metadata is a demanding task that is both complex and time-consuming (Michener, 2006; Pryor, 2007). Within the lifecycle, metadata can be created at virtually any point: prior to data creation, when files are saved, or when submitting to a repository (Pryor, 2007).

Metadata is ephemeral (Gray, Szalay, Thakar, Stoughton, & vandenBerg, 2002). It can be very difficult for researchers to find either the data or any contextual metadata as projects are completed. As funding ends, graduate students and staff who may have created and managed the data leave. With them goes all of the knowledge of the data (Pritchard, Anand, & Carver, 2005). Even data sets with rudimentary metadata are difficult to find and, thus, are lost to the community (Swan & Brown, 2008).

Metadata is both for human use, as in search and discovery, and for automatic processing (Buneman, Abiteboul, Szalay, & Hagehülsmann, 2006). The metadata needed for automated preservation activities is difficult to create manually. Many of the tools being developed to help scientists document their data focus on data production and publishing more than on preservation (Borgman, Wallis, Mayernik, & Pepe, 2007; Cannataro, Congiusta, Pugliese, Talia, & Trunfio, 2004; Cheung, Hunter, Lashtabeg, & Drennan, 2008; Chin & Lansing, 2004; Frew & Bose, 2001; Holmes, Johnson, & Miller, 2004). Within some domains, automated tools for creating provenance or lineage and file level technical data have been developed as well as some tools for creating scientific metadata, including description of experiments, treatments, participant responses, data cleaning efforts, and information about the data creators (Cheung et al., 2008;

Chin & Lansing, 2004; Myers, Allison, Bittner, Didier, Frenklach, Green, et al., 2005; Myers, Pancerella, Lansing, Schuchardt, & Didier, 2003; Simmhan, Plale, & Gannon, 2005). Self-describing data and the software to process that data are not yet realities (Bose & Frew, 2005).

2.5.3 Preservation Technologies

The data collections model as developed by the National Science Board (discussed above) does not consider the technology required to preserve the data. Creating a persistent data collection requires a set of technologies that Moore (2007) describes as the preservation infrastructure. Central to any preservation infrastructure is a physical repository. Through much of the literature on the technology infrastructure of scientific data, the term "repository" often refers to a simple data store for datasets (Venugopal, Buyya, & Ramamohanarao, 2006). A broader view defines a repository as both a system and set of services designed as an archive for digital data with context, fixity, and persistence. Repositories provide services to ensure the long term archiving of and continuing access to data including backups, contingency planning, process resumption planning, hardware and network redundancy, automatic failover, and site mirroring (Kowalczyk, 2008). Repository services increase in importance as the amount of data grows (Choudhary, Kandemir, No, Memik, Shen, Liao, et al., 2000). "Preservation technologies" refers to all of the technical infrastructures to support data collections.

When institutions create repositories, they do so with numerous goals. These often include grant fulfillment, peer review, long term archiving and preservation, daily scientific practice, the leveraging of scarce or endangered resources, and exploiting large infrastructures (Kowalczyk & Shankar, 2011). However, the adoption of institutional repositories as an infrastructure for storing and disseminating scientific data is not widespread (Lyon, 2007). In the U.K., the Joint

Information Systems Committee (JISC[1]) contends that looking after scientific data is a key strategic challenge for repository administrators that will involve changes in organizational structures and cultures (Key Perspectives, 2010).

For data repositories, storage infrastructure is an important component to archiving and preserving data. Although storage has become an increasingly cheaper commodity (Atkins, 2003; ARL, 2006; Hacker & Wheeler, 2007), carefully designing an infrastructure is important (Rajasekar, Marciano, & Moore, 1999; Brown, 2003). Moore, Baru, Rajasekar, Ludaescher, Marciano, Wan, 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 option (Moore et al., 2000a; Moore et al., 2000b; Morris & Truskowski, 2003). Criteria used to assess scalability, expandability, and extensibility of storage technologies include longevity, capacity, viability, obsolescence, cost, and susceptibility to failure. These criteria can be used to develop a matrix to measure the media usability both in terms of life span and 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).

For most scientists, such preservation infrastructure lies beyond their reach. They have neither access to a repository (Lyon, 2007) nor expertise to build a scalable storage infrastructure (Henty et al., 2008). Those who do have access to a repository find complicated workflows and difficult metadata creation processes limit both their ability and incentives to deposit (Crow, 2002; Lyon, 2007).

[1] The U.K. Joint Information Systems Committee (JISC) supports higher education institutions in the U.K. providing strategic technology services and funding for research.

2.6 Antecedents as Barriers to Preservation

The antecedents to preservation – data management, contextual metadata, and preservation technologies – can also be barriers that prevent preservation. Rusbridge (2007) describes both the positive and negative navigation of these barriers. For projects with stable staffing and good communication, good sense can be sufficient to manage the data well enough to produce sound scientific results. But many projects produce data that is both unknowable and unusable – that is, without context, without the associated experimental conditions, in undocumented files, and in incomprehensible spreadsheets (Rusbridge, 2007). It is this second scenario – unidentifiable or unusable data – that is the primary barrier to preservation. Data that cannot survive the short term certainly cannot be preserved (Lynch, 2008).

2.6.1 Data Management as Barrier

As discussed above, data management is a crucial antecedent to preservation, but it is also a crucial barrier. For scientists, data management can be a low priority, can require skills and expertise not readily available, and can cost more than its perceived value. Although data preservation is important to funding agencies, most researchers are rightly focused on their science. With all academic incentives rewarding new work, it is counterintuitive for scientists to invest time, effort, and money to care for older data. Data management is frequently considered to be overhead and not research (Anderson, 2004) and can be a burden for scientists (Bell, Hey, & Szalay, 2009).

Data management is a technical skill that requires an understanding of storage technologies, data replication strategies, and contingency planning. Researchers often lack the skills to be effective data managers (Treloar, Groenewegen, & Harboe-Ree, 2007). Anderson (2004) contends that data managers need domain specific knowledge in order to understand how

best to manage the data. Ideally, all stakeholders need to be involved in providing requirements for data management. But the self-sufficient research culture (Pryor, 2007) often hinders collaboration between scientists and trained data managers (Committee on Data for Science and Technology, 2002). Rather than hiring data management experts, scientists have been using Ph.D. students as systems administrators, sacrificing a generation of new researchers (Hey, 2010).

Data management is expensive in terms of both personnel and equipment. Frequently, scientists lack the necessary funding which would allow them to develop a robust data management infrastructure (ARL, 2006). So in the absence of data management as a project is in progress, data is too frequently abandoned, transferring any data recovery costs to the future with significant risks of both loss of data and loss of context (Lord & Macdonald, 2003).

2.6.2 Metadata as Barrier

Creating metadata, that is making the data intelligible, can be a major impediment to preservation (Lyon, 2007). Data repository managers have developed guidelines that promote good metadata practices. These practices include documenting data throughout the research project and creating an audit trail of all of the data processing transformations wrought over the data life cycle (Vardigan, Heus, & Thomas, 2008). However, the process of creating metadata, what is described as the "mechanics of metadata," causes both confusion and frustration among researchers (Swan & Brown, 2008, p. 16). Cheung and colleagues (2008) state that preservation is stymied by a "lack of simple tools for publishing data with provenance information, lack of motivation for scientists to spend time and effort preparing their data for publication, concern with intellectual property rights, lack of standards for publishing datasets and discipline specific

tools" (p. 5). The incentives for creating usable, machine-processable metadata are not strong enough to overcome this absence of useable tools.

2.6.3 Preservation Technology as Barrier

Data stewardship needs to be a shared responsibility. The researcher is initially responsible for data, but responsibility needs to be transferred to an institution for long term archiving and preservation (Lynch, 2008). The infrastructure that institutions have developed to help share the burden of data preservation can present barriers that make fulfilling their mission difficult.

The major barrier for preservation is the complicated, inflexible, and counterintuitive processes required to deposit data within these repositories (van Westrienen & Lynch, 2005). Steinhart (2007), describing the use of an institutional data repository, states that even low barriers to a technology might not be low enough for researchers to use. In an international survey of institutional repositories, van Westrienen and Lynch (2005) found that a vast majority of data submission was the work of librarians, not the researchers. The literature is sparse on specific examples of data submission issues. Two apparently successful deposit processes both emphasize multiple submission methods for data: web forms, batch upload, and spreadsheet to metadata standard conversions (Kandasamy, Keerthikumar, Goel, Mathivanan, Patankar, Shafreen, et al., 2009; Barrett, Troup, Wilhite, Ledoux, Rudnev, Evangelista, et al., 2009).

The repository culture may be negatively influencing preservation. Much of the research in digital preservation has focused on building the repository. This focus on repositories has created a model of preservation that is post hoc in that the repository tries to gather as much information as possible after the object is created and before it is ingested into the repository.

However, these efforts are often too late because data is missing; data is not discoverable; data is not recoverable (Kowalczyk, 2008).

2.7 Gaps in the Literature

The literature on digital preservation spans a number of disciplines – information science, library science, computer science, as well as a number of scientific domains such as biology, astronomy, and environmental science. This rich body of literature has multiple research streams, many of which are becoming better developed. Yet, gaps remain.

As discussed in section 2.2, the literature on data lifecycles has some significant gaps. The lifecycles are either completely generic (Green, 2008; Green & Gutmann, 2007; Martinez, 2008; Rumsey, 2010) or based on a very narrow domain (Borgman, 2007; Borgman, Wallis, & Enyedy, 2006; Higgins, 2008; Karasti, Baker, & Halkola, 2006; Wallis et al., 2008). These lifecycles do include data management, data preservation, or data archiving, but these phases are generically vague. These life cycles leave a number of questions unanswered: What actually is preserved – all files, some files, or final files? Where will this data be preserved/archived – on lab storage devices, in institutional repositories, or in community-supported research collections? Who will be the long-term caretakers or curators of this data – the individual researcher, the institution at which the research was conducted, or the funding organization? These questions are a symptom of a larger problem: the paucity of theory to explain and model the preservation of scientific data.

Defining and categorizing the threats to preservation has been a significant theme in the literature. Most of these threats are discussed within the context of a repository. More research needs to be done in order to understand how these threats affect data throughout its lifecycle.

Determining the lifecycle stage at which each threat is most probable could be used to develop policies to mitigate these threats and to promote preservation.

There is surprisingly little quantitative data describing the perceptions and behaviors of scientists in the digital preservation literature. Much of the quantitative data used in the literature was collected as part of a project to develop systems, services, and policies for data preservation within a single organization or a consortium (Henty et al., 2008; Jones, Ball, & Ekmekcioglu, 2008; Lyon, 2007; Marcus et al., 2007; Pritchard, Anand, & Carver, 2005; Pryor, 2007; Witt, Carlson, Brandt, & Cragin, 2009). Little of this data was used to develop theory. Theory development backed by quantitative data is necessary for the field to progress.

2.8 Research Questions

Archiving and preservation of scientific data can no longer be thought of as a post-project activity (Anderson, 2004). Preserving digital data should be an important function of scientific infrastructures (Hacker & Wheeler, 2007). However, a there is a lack of "evidence from the community of active researchers with respect to their own needs and aspiration within the research life cycle" about data management roles and responsibilities (Pryor & Donnelly, 2009, p.167). New research is needed to describe, measure, and mitigate the "obstacles to the longevity of digital materials" (Ross, 2007, p. 6).

This dissertation proposes to illuminate some of the antecedents to preservation by describing the environment in which digital scientific data exists. This paper will propose a model of the scientific data environment that begins to capture the complex interactions among the data, the environment, and the community. The research questions that frame this dissertation are as follows:

1. What are the data practices of researchers and scientists?
2. How can the research practices be reflected in a lifecycle for research data?
3. Are the antecedents to preservation actually barriers?
4. How do the threats to preservation affect the data lifecycle?

The research was conducted in two phases: 1) case studies of scientific laboratories and centers for grounded theory development and 2) a survey of scientists to quantify the antecedents to preservation.

3 Preliminary Study

To begin to develop a theoretical framework for better understanding data preservation and the environments in which data is created, a preliminary study was conducted (IU IRB Study Number 06-11593). The preliminary study used the grounded theory methodology to develop a model of the data preservation environment. This study used a series of questions to facilitate semi-structured conversations with the directors of 11 research centers and laboratories in a variety of domains from three different universities. These questions fall into the general areas of interest to be studied: the amount and types of data to be preserved, the technology in which the data exists, data quality, and scientists' perceived need for data preservation. This section describes the preliminary study, the initial results, and gaps that need further inquiry.

3.1 Methodology

3.1.1 Grounded Theory

Grounded theory is a qualitative research methodology for inductively deriving theory based on the data gathered about one or more phenomena (Strauss & Corbin, 1990). Grounded theory provides a framework and a well established set of procedures to discern patterns in the data; the framework includes a set of systematic and flexible guidelines for both collecting and analyzing qualitative data as well as for constructing theories that are grounded in that data (Charmaz, 2006; Urquhart & Fernández, 2006). In other words, grounded theory is used to develop concise theory from the rich qualitative data generated from interviews.

The grounded theory methodology is particularly useful for emerging areas of study (Eisenhardt, 1989; Sarker, Lau, & Sahay, 2001). Friedman (2003) posits that building theory

allows researchers to move beyond a succession of unique cases to broad explanatory principles that can help to solve many kinds of problems. As an emerging area of study, digital preservation and data curation can certainly benefit from a theoretical framework grounded in the data.

3.1.2 Research Sample

The process of determining the research sample for the grounded theory methodology differs from the process of determining the population in hypothesis-testing research. In grounded theory, theoretical sampling is preferred to random sampling (Glaser & Strauss, 1967; Eisenhardt, 1989). Theoretical sampling allows researchers to choose cases that replicate or extend theory. By determining theoretically relevant categories and choosing cases *a priori*, researchers can create a diverse set of participants. Cases can be selected based on a number of criteria: the typical or representative case, the negative or disconfirming case, the exceptional or discrepant case, and polar types (Miles & Huberman, 1994; Glaser & Strauss, 1967; Voss, Tsikriktsis, & Frohlich, 2002). During the course of the research, cases can be added or eliminated as the research questions or frameworks are extended. Glaser and Strauss (1967), the creators of the grounded theory method, contend that theoretical sampling provides researchers with multiple options for gathering data that includes different views or vantage points from which to understand a category and to develop its properties.

Participants for this study were chosen based on the theoretical sampling model of polar examples – that is, participants that are diametrically opposite extremes. Eleven research centers and laboratories from three different universities were chosen based on four theoretically significant categories: size of lab, funding, scientific domain, and type of science (see Table 3).

Table 3. Sample Description

Theoretical Sampling Categories	Number of Participating Centers and Laboratories (11 total)
Large Lab (5 or more researchers)	7
Small Lab (<5 researchers)	4
Well funded	6
Poorly funded	5
"Big Science"	6
"Little Science"	5
Physical Science Domains	4
Biological/Medical Science Domains	4
Informatics Science Domains	3

The sample contains both large and small labs. A recent JISC report indicates that size of lab can have an effect on data curation; larger labs have more resources to manage their data (Key Perspectives, 2010). Although no specific definitions of large or small labs exist, this study defines a large lab as five or more researchers while a small lab has fewer than five researchers.

Lab funding can have similar impacts as lab size on preservation: more funding means more resources to apply to curation activities. As well, outside funding agencies can influence preservation by mandating data management policies or repository deposit (Coles, Carr, & Frey,

2007). The sample contains both well funded and poorly funded labs. For this study, a well funded lab is defined as one with both base funding and a sufficiently constant grant stream to keep the researchers for multiple years and across projects. A poorly funded lab is defined as one without base funding or without a steady stream of grant funding.

Type of science, the third category, categorizes scientific research as "big science" and "little science" (Weinberg, 1961). For this study, "big science" refers to the work of any laboratory that requires massive capital investment to yield results; "little science" refers to the work of any laboratory that does not require significant capital investment. Science type has been widely used in the data preservation literature to categorize data standards, data creation mechanisms, volume of data created, and funding (Borgman, Wallis, & Enyedy, 2006; Key Perspectives, 2010; Lyon, 2007; Wallis et al., 2008).

Scientific domain is another theoretically important distinguishing category. As with type of science, domain is used through the literature. Data practices vary widely between scientific domains (Wallis et al., 2008). Technical standards for both the data and the metadata as well as data storage standards are well established in some domains while nonexistent for others (Key Perspectives, 2010; Lord & Macdonald, 2003; Lyon, 2007). The sample for this study used a number of scientific domains, including biological/medical sciences, physical sciences, and informatics-based sciences.

3.1.3 Data Collection and Analysis

The data for this study was collected in one-hour semi-structured interviews with the directors of 11 laboratories and research centers at three universities. The questions were used as an introduction for a conversation (see Table 4). Following guidelines for grounded theory research, the interviews were not taped, but extensive notes were taken during and immediately

after each interview. Glaser and Strauss (1967) strongly advise researchers to use field notes rather than recorded interviews, explaining that too much data obscures the essential information that would naturally be retained in the researchers' memories. Stern (2007) contends that recent methodologists' emphasis on complete word-for-word accuracy focuses the research on rich description rather than on theory. Researchers "need to focus on the accuracy of their discovered truth, rather than the less important what-did-they-say-exactly" (p. 119).

In grounded theory, analysis and data collection are ongoing, recursive, iterative activities: data is collected; the data analysis begins; more data is gathered and analyzed. The data for this study was coded iteratively, starting with an open coding mode to develop the first level concepts, followed by selective and theoretical coding. The coding results in categories, which are the conceptual elements of the theory (Dey, 2007). The categories established during the planning phase of this research (as listed in Table 4) changed significantly as the data was collected and analyzed. The final categories that emerged from the data are data creation, quality control, content, format, context, data collections, and technical infrastructure. Each of these categories has a number of properties, which are the conceptual aspect of the category (Glaser & Strauss, 1967). A full explication of these categories, their properties, and their relationships follows.

Table 4. Survey Questions

Initial Category	Questions
The Nature of the Data	• Please describe the types of data that you use in your research – size of files, file formats, uniqueness of your data.
Data Priorities	• What data is the most important to your research? Why?

	• Do you have a formalized set of criteria for judging the quality of your data? Can you provide the criteria?
Preservation Awareness	• Do you worry about the longevity of your digital files? Please name your concerns. • What data do you think is most at risk? Why?
Scale	• Can you estimate the amount of data that you would like to archive over the next 2 years? 5 years? 10 years? • How many people work in your lab? On your specific research?
Solution Options	• What solutions can you suggest for preserving your data? • Within the context of your main application, would having the ability to create a trusted digital object be worthwhile? Why?

3.2 An Emerging Model of the Data Environment

In this section, data collected from the discussions with the directors of the 11 scientific research centers and laboratories described above was used to develop a theoretical model of the e-Science Data Environment describing the antecedents to preservation. The e-Science Data Environment is the socio-technical situation in which scientists create, use, and store their data. The data environment is a complex interaction of content, formats, context, quality control, data collections, and the technical infrastructure of the researcher's home institution (see Figure 5). Each of these constructs is described and discussed below.

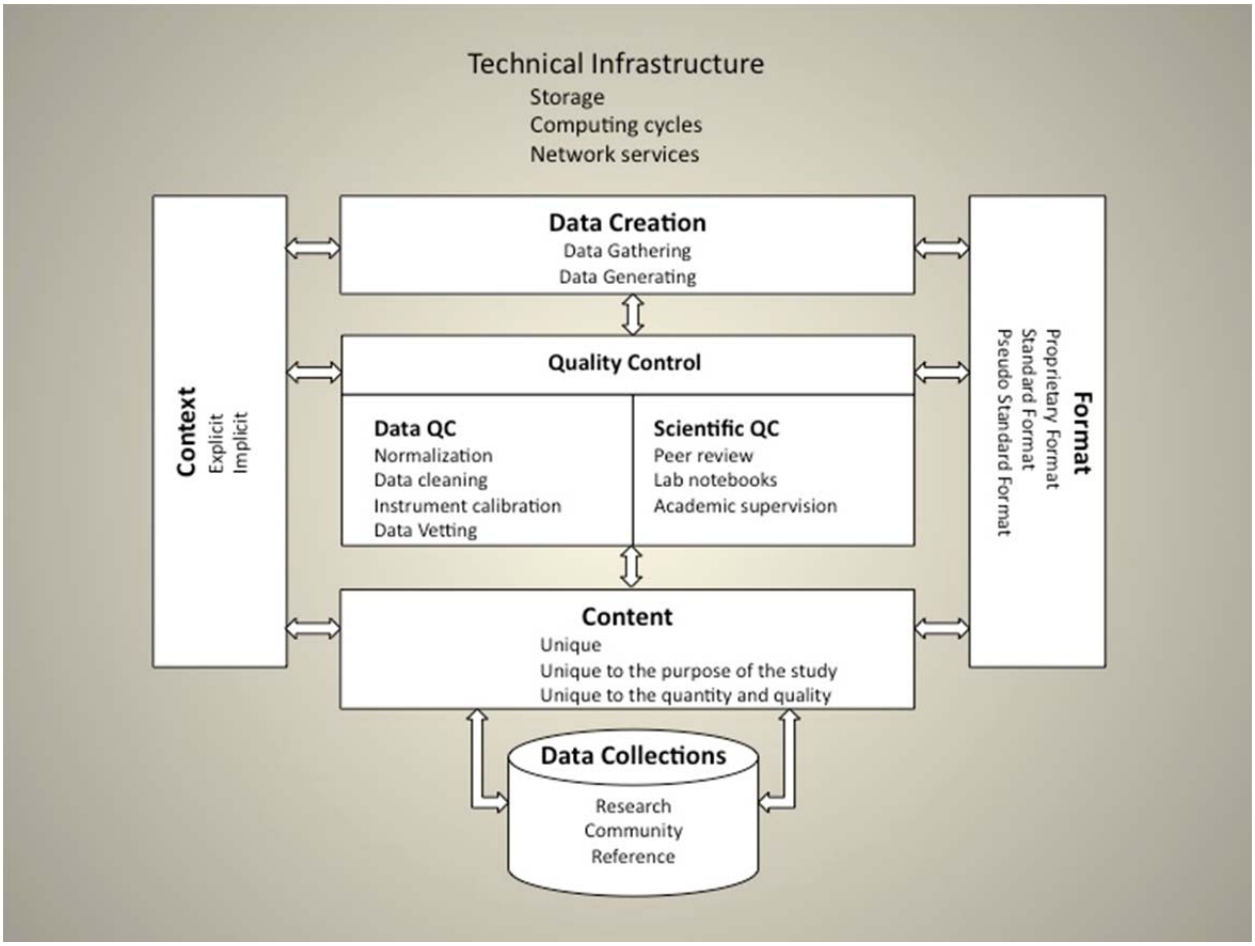


Figure 5. e-Science Data Environment

3.3 From Data to Content

Scientific research data can be generated though an experiment or observed via instrumentation; or data can be gathered from existing sources such as government data, vendor data, or web crawls. None of the scientists interviewed was exclusively a creator or a gatherer. Regardless of source, most of the scientists interviewed processed their data by merging data, interpreting and mapping multiple metadata formats, or integrating data with different levels of precision and scale. Data becomes content as value is added via quality control processes, format conversions, contextual data, and structural metadata.

Assessment, the process of determining preservation priorities, is a thread through the digital preservation literature. The criteria for assessment generally include a binary judgment of uniqueness: data is unique and should be preserved; or data is derived, can be recreated, and should not be preserved (Gray et al., 2002; Henty et al., 2008; Key Perspectives, 2010; Lord & McDonald, 2003; Lyon, 2007). The results of this study show that uniqueness is more complicated than previously thought.

From the interviews, three levels of uniqueness emerged. The first is the truly unique - no other holdings of this data exist: the preservation-worthy data as described in the literature. This type of data is often the result of experiments or observations. The second level is unique to the purpose of the study: millions of slides of mouse livers exist, but none have this specific treatment for this specific research question; or data derived from external sources such as reference collections with this unique analysis. The third level is unique because of the quantity and quality of the data: that is, the level of uniformity and integration of the data, the breadth of data, the longitudinal nature of the data, or the added value of metadata. Throughout the literature, the processing of data to create this uniformity or integration is often characterized as simple computation (Gray et al., 2002; Henty et al., 2008; Key Perspectives, 2010; Lord & McDonald, 2003; Lyon, 2007). But to these scientists, the process is very costly in terms of staff, equipment, time, and intellectual effort. It is the processing, both manual and automated, that is unique; thus, all data ultimately becomes unique. It should not be inferred that the scientists wanted all of their data preserved, but it does indicate that uniqueness cannot be the sole assessment criteria.

The scientists, themselves, struggled to assess the preservation priority of their own data. Final, publication data was easily to prioritize; but analysis data and intermediate processing data

were very problematic. One scientist summed up the concerns of many, stating that it was hard to know what will be important in the future.

3.4 Quality Control Processes

The scientists had two very distinct understandings of quality: the quality of their data and the quality of their science. In addition to ensuring that the original data is correct, quality of the data included processes such as normalizing the data to allow accurate merges from disparate sources. The scientific process has a well established quality control mechanism in peer review. This process includes vetting by peers, meticulous and irrefutable record keeping via lab notebooks, and academic supervision of students. But there is growing concern that peer review is failing to judge adequately the quality and the scientific value of the content of datasets (IWGDD, 2009; Riley, 2006; Swan & Brown, 2008).

The discussion of a formalized set of criteria for judging the quality of data elicited more equivocal and angst-filled conversation than any other. Many of the researchers responded with a "yes and no" answer – yes they have some criteria, but no it is not formalized or codified. One researcher did not see a need for quality control stating that "the data is the data," but all of the others saw quality control as a primary imperative of their directorships. One director with a large lab staffed with doctoral students and post docs described the quality control situation as a "bit of the Wild West here." He wanted to establish more control of the process, but he was concerned that rigid standards would compromise creativity. One researcher seemed to sum up all of the others' angst, stating that he did not have a set of formalized quality control criteria, "but I wish I did!"

The processes that the scientists use to ensure quality of the data depend on the type of data and the source of the data. For original data generated from equipment, the hardware provides a significant level of quality control. The quality of the data depends on the quality of the maintenance, testing, and calibration of the equipment. For data gathered from existing sources such as vendors or web crawls, data needs to be manipulated, merged, normalized, reconciled, and cleaned.[2] These processes can introduce errors and require a community effort to ensure quality control as described below.

Many of the scientists saw their data in two distinct ways – as intermediate process data and production data. Some of the scientists saw the intermediate data as important while others saw it as a byproduct of their main work. Many of the researchers expressed concern over this process data. A genomics researcher relayed a story that shows some of the issues. Process data that was seen as an unimportant byproduct of the production data was not saved but is now necessary to interpret the quality of the production data. Since it is lost, it must be recreated at great expense in both time and money. The scientists were unable and unwilling to predict which intermediate process data would be important for the future.

Many of the scientists were interested in building and maintaining their data over a very long time. For some, this meant decades of data and for others, this meant centuries of data. Karasti, Baker, and Halkola (2006) posit that maintaining the coherence and continuity of longitudinal data requires consistency in documentation and maintenance of the digital archive. This is the grand challenge to the scientists interviewed. They were very concerned about consistency of the data and the adequacy of their metadata. Several of these lab directors are

[2] Data cleaning is an imprecise term meaning to fix errors such as spelling and punctuation, to remove duplicate data (commonly known as "de-duping"), removing spaces and other extraneous characters, and other similar operations.

having difficulties finding older data to fill in the gaps. One scientist told of recently discovering some old VAX tape filled with data that was thought to be lost. Others told of creating digital data from accounts in books or from old notebooks. All of the scientists who work with longitudinal data are concerned about keeping it viable for the future. Several of the directors with significant longitudinal data discussed data management as a quality control issue. Having multiple copies of their data in multiple locations was a part of the quality control process.

Data integrity, the expectation of data quality, is a significant outcome of quality control processes. Rigorous quality control is a link in the chain of trust assuring that the data is "whole" or complete, consistent, and correct. Building trust in the data and creating data with integrity are important steps in the scientific process; thus, quality control is an important component in the e-Science Data Environment.

3.5 Data Collections

In the e-Science Data Environment, data collections represent the disposition of the data when the research is complete. As described in section 2.1, data collections are defined as the infrastructure, organizations, and individuals needed to provide persistent access to stored data (NSB, 2005). This model uses the three-layer typology of data collections of 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 use the data standards of its community. Community data collections generally serve a well defined area 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 (NSB, 2005).

None of the researchers in this study would have used the NSB typology, but their data fits nicely into the data collection scheme. All of the researchers had large research collections. These individual collections were of the most concern to them. A number of the researchers had contributed data to either community or reference collections. One of the researchers was actively involved in developing a community collection and expressed his significant concerns about the long-term sustainability of this collection; he was unsure if the community would be able to continue to provide sufficient funding for the ongoing maintenance, technology upgrades, and storage requirements.

3.6 Context

Context describes the relationships of the data content to its environment (Consultative Committee for Space Data Systems [CCSDS], 2002). Data is situated in a context that includes how data fits into the physical and technical environments (file formats and field descriptors) as well as into the scientific environment (experiment treatments and applications) (Kowalczyk & Shankar, 2011). As discussed in section 2, contextual metadata is both a vital component of and a major barrier to preservation.

One of the scientists interviewed in this study considered context to be the knowledge about the data. As a knowledge object, this context is a complicated interchange of implicit and explicit metadata (Kowalczyk & Shankar, 2011). Contextual data can be explicit, tangibly and specifically expressed, or implicit, implied by relationships. In this study, explicit context consists of lab notebooks, data stores in excel or databases, or metadata in community standard

formats as discussed above. But much of the contextual data is implicit, implied in file organization structures or in file naming schemes.

The scientists in the study were deeply concerned about the lack of context for their data. Of greatest concern is removable storage media without labels, which virtually orphan the files that several of the scientists had considered archived. One scientist was deeply concerned about the relationships among the time-sequenced images from his experiments and confessed that he was deeply relieved when the paper was published, as he then considered that contextual data to be archived.

The contextual data fell into four categories: 1) data about the experiments, 2) the relationship among files, 3) data quality control algorithms or software, and 4) social data such as usage statistics or discussion forum data where feedback about the data is gathered. Experimental contextual data is generally explicit stored in lab notebooks, spreadsheets, and automated workflows. This data is crucial to the quality of the science and, thus, to all of the scientists surveyed. The scientists were committed to maintaining this data and were confident about the accuracy of the data. Structural context such as the relationships among files was more problematic. Much of this structural context is implicit in file names and directory structures, but most of the explicit data is often stored in opaque application files. Structural context can be an important component of the science because it can document the relationships of files in time sequence over the course of the experiment or spatial relationships of virtual slices of an image. Software was identified by several of the scientist as being very important for their data. These scientists were concerned about the ability to use their data without the software that created it, the software that did the analysis, and the software that rendered or visualized the data. Increasing in importance to the scientists is social data that exists in web usage logs, blog entries,

or listserv environments. While this data is primarily explicit, much of the data is free text, thus so unstructured as to be difficult to use.

Within this model, context is more inclusive than metadata. Metadata is codified information about data, generally using one or more predetermined structured representational format. This static data can be considered one aspect of context. But context is also dynamic (Aktaş, Fox, & Pierce, 2005), relative, interpretive, and imperfect (Soylu, De Causmaecker, & Desmet, 2009). That is, context describes a variable and volatile situation that, when represented as metadata, can be inadequate and incomplete.

Contextual information is generated and accumulated at every step in the model (see Table 5). In data creation, whether data that is gathered or generated, contextual data includes information about provenance – for example, data sources, instrumentation settings, and experimental variables. As data is processed and becomes content, data quality issues, purpose, identification, definition, and description data are added to the contextual data. During data quality control processes, information about the normalization, data cleaning, and integration processes become part of the contextual data. When the data is published or archived, additional contextual information is generated: contributors, date published, use restrictions, funding partners, publications resulting from the data collection.

Table 5. Contextual Data

Model Component	**Contextual Data**
Data Creation	Data sources Instrumentation settings Experimental variables
Content	Data quality rating

	Purpose
	Identification
	Definitions
	Descriptions
Quality Control	Data cleaning/integration algorithms
	Data normalizing algorithms
	Processing software
Data Collection	Contributors
	Date deposited
	Access and use restrictions
	Funding partners
	Publications
	Ongoing data annotations

Traditionally, creating this contextual metadata has been a one-time event, usually at the very end of a project. However, new social networking paradigms of user-contributed metadata are being applied to data; collections are beginning to ask subsequent data users to add additional information via textual annotations (Chin & Lansing, 2004; Ives, Halevy, Mork, & Tatarinov, 2004; Myers et al., 2005; Michener, Beach, Bowers, Downey, Jones, Ludäscher, et al., 2005; Jaiswal, Giles, Mitra, & Wang, 2006) 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 itself, 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, 2006).

3.7 Formats

Digital data is represented in a file format, which is defined as the internal structure and encoding that facilitate computational processing as well as rendering for human use (Brown, 2006). Abrams (2004) contends that the concept of representation format is the foundation of many, if not all, digital preservation activities. Pearson (2007) states that format change and/or obsolescence is the major threat to preservation. The proliferation of complex formats greatly increases the complexity of preservation (Barateiro et al., 2008; Ross, 2007). The technical format of a file affects its probability of being preserved (Kowalczyk, 2008).

The initial expectation was that scientific data formats would easily fall into two categories: 1) proprietary, defined as a format owned by one or more organizations or individuals with legal restrictions of use, with limited transparency and/or software for rendering and processing and 2) standard, defined as a format in the public domain or owned by a organization that makes the format available with no legal restrictions and has publically available documentation and software for processing and/or rendering. Proprietary formats used for instrumentation data, internal systems data, or vendor data are often migrated to standard formats. Standard formats used by this set of scientist were image standards such as TIFF, JPEG, or community XML standards like FITS, BSML, and FGDC.

More than half of the scientists surveyed confused domain data standards with generalized computing data standards; thus, a third category emerged: pseudo-standards. A pseudo-standard is a generic syntactic data computing or storage format without any semantics

such as such as CSV, ASCII files, and SQL/Xpath databases. The scientists declared that they used standard formats and named these pseudo-standards as their formats.

Format affects every aspect of the e-Science Data Environment model because format is the basis of every data file. Format determines the syntax and often the semantics of the data. At data creation, the format may be determined by the mode of collection. Vendors may prescribe the format for both data created by instruments and data gathered from databases, websites, or applications; data could be created in a community standard; or data could be created in a format created specifically for this particular study or for the individual laboratory. The format of the original data affects efforts to integrate data from different sources and for data quality processing such as normalization and data cleaning. The resulting content must be stored in a format. For research collections, format is the choice of the individual researcher or laboratory. But community and reference collections mandate specific formats. It is the responsibility of the researcher to conform to the format of the collection prior to submission. Format choices earlier in the data environment can affect the degree of difficulty in the migration to the collection format requirements.

3.8 Technical Infrastructure

The most surprising finding of this preliminary study was the importance of the technical infrastructure. This is the core technology of the researcher's institution that includes data storage as well as network and computing resources and that underlies the entire model. It was, by far, the largest factor in deciding how scientists dealt with their data. The scientists in this study were located in three different universities. The large midwestern state university has a large computing and storage cloud with no direct cost to researchers. The medium midwestern

state university has a large computing and storage cloud, but costs are allocated by usage to researchers. The Ivy League university provided only the network backbone and required researchers to create their entire technology environment including storage, computing cycles, physical space, electrical infrastructure, and personnel such as systems administrators and database administrators. When high quality storage was available at no cost, more data was stored in larger, standard formats. When high quality storage was a chargeback to the project or if the researcher had to create the storage infrastructure, data was stored in the most economical format, often on removable media such as CD-ROMs.

For virtually all of the scientists, preservation was equated with better hardware. They perceived faster computers with larger capacity hard drives to be the solution to their problems. When pressed, they could not see how hardware would help them deal with the contextual data that is so important to them. They expressed concerns about their dependence on specific vendors for both hardware and software. One scientist has significant concerns about the necessity of using proprietary GIS software, particularly issues with format change, migration, and backward compatibility for consumers of the data. Two researchers indicated that their labs used only open source software or developed their own applications at significant costs to avoid becoming dependent on vended systems. One of these researchers felt that this effort took time away from her science: time to manage the software development process and time necessary to secure funding via grants. Proprietary storage technologies were concerns for other scientists: VAX formats and old magnetic tapes were cited as examples.

3.9 Preservation Awareness

A key element of this study was to discover the preservation perceptions of the directors of scientific laboratories. The scientists surveyed expressed deep concern about the longevity of their data, both the actual data and the contextual metadata. Several of the scientists have contractual obligations to keep data at least 10 years. Interestingly, none of the labs reported requirements for deleting research data. The scientists are concerned that they will not be able to keep the data useable for that duration. There were anecdotal stories of data loss, some of which was data important for their longitudinal studies. Researchers from communities like astronomy with reputations for good data preservation and data management have concerns with the usefulness of their preservation infrastructures. One scientist described a community "write only archive" in which researchers can deposit data. This archive does not accommodate metadata and has no "scientist friendly tools" for data retrieval. To restore some lost data, he had to pay the archive for staff time. The process took over six weeks to get the raw data with an additional two weeks to reprocess the data. He questions whether this is really a preservation service.

3.9.1 Preserving Content

There is growing consensus that preservation needs to be included in the entire lifecycle of data. Lavoie and Dempsey (2004) contend that digital preservation techniques are most effective when they are pre-emptive. In most circumstances, digital preservation needs to be considered at data creation – at the application level when the data needed for preservation is most readily available (ARL, 2006; Atkins, 2003; Galloway, 2004). To understand better the application space in which scientists work, the scientists were asked about the application space for their domain and whether they would want a function to be able to create a preservable object, that is a complete archival package that could be ingested into a preservation repository.

Overall, the support for "built in" preservation was positive (73%). The level of support varied depending on the nature of the application space. Scientific software applications are specific to a domain or community (such as BLAST, a set of genomics tools) or specific to the problem (such as meteorology modeling). Some domains use general applications such as GIS, statistical, or mathematical tools. Two of the research centers in the study were microscopy laboratories whose scientists used no software external to their instrumentation. For scientists with a community-sponsored application, the support for integrating preservation into the application was strong (see Table 6). For scientists with a vended or general application, there was moderate support. For scientists with applications for a specific problem, the support was mixed. For scientists who used no external application, there was no perceived need for preservation services.

Of the three researchers (27%) who said that they did not think this would be useful to them, none worked in an environment that had a "main software application" for doing their work. One of these researchers served government data to local researchers and felt that the responsibility was with the others who used the data. Two of these researchers were directors of microscopy labs. Although one had expressed concern about preserving the data, he had also participated in an experimental project to preserve a set of images. The work required to capture the metadata was too expensive and time consuming. He could not envision a simpler solution than the one he had experienced.

Table 6. Archiving within an Application

Application	Level of Support
Math, stats, or other vended packages	Moderate support

Community sponsored applications	Strong support
Individual application for specific problem	Mixed support
No application	No need perceived

3.9.2 Funding and preservation

Funding was a significant factor in the probability that these scientists would be able to overcome the barriers to preservation. Several of the very small labs were funded solely on soft money, that is, non-renewable grant funding. One of the directors described his lab as a "shoe-string operation." Having insufficient funds causes the labs to choose the least expensive storage option rather than the one most reliable or efficient for preservation. These labs had no resources for documenting the context of their data or for ongoing data management. Even labs with base funding in addition to their grant funding had insufficient funding for long-term data management. The other major funding issue was end-of-project issues. When well funded projects run their cycle, there is no provision for preserving the data that was produced in that project. Not only was funding for preservation of the data not included in the grant, but no money from any of the centers' operational budget was allocated. As discussed previously, the amount of technical infrastructure provided to a lab at low or no cost is the biggest factor in its ability to overcome the barriers to preservation.

3.10 Evaluating Newly Developed Grounded Theory

Urquhart (2007) argues that although grounded theory has been used in information systems research frequently, it has not actually generated theory. Rather, it has been used to develop descriptions of phenomena. Weber (2003) contends descriptions, if they explain or predict some phenomena, can be considered theory. Eisenhardt (1989) defines good theory as

parsimonious, testable, logically coherent, and grounded in data. It should demonstrate sufficient evidence to justify its conclusions. It must have a good fit with the data. It should present "new, perhaps framebreaking, insights" **(**Eisenhardt, 1989, p. 548). Strauss and Corbin (1990) set four general criteria for evaluating newly developed grounded theory: fit (the theory should have fidelity to the reality of the area of interest); understanding (the theory should be "comprehensible and make sense"); generality (the theory should be comprehensive, abstract, and conceptual with broad applicability); and control (the theory should provide a "meaningful guide to action") (p. 23).

The e-Science Data Environment model (see Figure 5 above) was developed using grounded theory methodology. Using the Eisenhardt (1989) and Strauss and Corbin (1990) criteria, this nascent theory may be considered "good theory." The e-Science Data Environment model fits the data by creating a faithful depiction of the process of creating and managing scientific data. Numerous stories and direct quotes from the study participants are evidence to support and justify the categories, the properties, and their relationships. The model presents a logically coherent and understandable view of the environment in which scientists create, use, and manage their data. By using a theoretically diverse sample, every attempt was made to create a generalized model encompassing multiple scientific domains. The model is parsimonious, using seven major constructs to describe the data environment that can be depicted in a single diagram. Others will evaluate the e-Science Data Environment model for its usefulness. Comments from external reviews have been encouraging.

3.11 Gaps in the Theoretical Model

Because the e-Science Data Environment model was based on a preliminary exploratory study, a number of gaps exist. This emerging theoretical model is promising, but it does not yet address fully all of the antecedents to preservation and needs to be tested, generalized, and enhanced.

Two large gaps are evident in the model. First, there is a lack of quantifiable data to describe the interactions among contextual metadata, format, and the phases of the lifecycle. It is not clear how often data is converted, how many formats researchers use, or what formats are used. The process by which metadata is created from contextual data is equally unclear; its timing in the processes, the resources required, and the formats used are yet unknown. The second major gap is the insufficient development of the impact of technology infrastructures on preservation. Although some data on data management and preservation technologies was collected, it was not of sufficient specificity to model effectively.

4 Developing a Generalized e-Science Data Environment

The e-Science Data Environment was developed using grounded theory methodology with interviews from 11 scientific laboratory or research center directors. To generalize the findings of this research quantitative data is required. Combining quantitative data with qualitative data can indicate relationships that neither type of data alone could reveal; the triangulation can substantiate the constructs (Eisenhardt, 1989). A survey instrument was developed to gather the quantitative data necessary to generalize and extend the model (Study # 1010002804; see Appendix A for approved forms). Constructs that were generalized include data collections, levels of uniqueness, and technical infrastructure. Constructs that were generalized and enhanced are quality control, context, and formats. In the following section, these constructs are operationalized as survey items.

4.1 Operationalizing the Research Questions

To generalize and extend the e-Science Data Environment model, the major constructs needed to be quantified. Operationalizing the e-Science Data Environment constructs, the process of creating measurable items from broad concepts, resulted in a survey instrument with five demographic/categorizing items and 31 substantive items. The survey instrument used a number of different quantitative question types: semantic differential scale, rating scale, agreement scale, multiple choice, and dichotomous. In addition to the quantitative data, this survey collected qualitative data by employing several open-ended questions. Open-ended questions were used in two circumstances: to allow respondents to explain their use of the Other/Comments option (i.e., when their circumstances are not described in the possible answer

choices) and to obtain data that has too many options to be listed for selection (e.g., the length of a contractual obligation or metadata formats used).

The following subsections describe each construct of the e-Science Data Environment model, develop new research questions to generalize and extend the model, and operationalize the construct.

4.1.1 Data Creation

Data creation is the simplest construct in the e-Science Data Environment. As discussed in section 3.1, data can be *generated* by observations, instruments, or experiments; or data can be *gathered* via databases, vendors, webcrawls, and other processes. Within the modes of data gathering and generation, scientific research data is created via a methodology, a strategy and a set of techniques for framing and solving research problems. Research methodologies have a deterministic effect on types of data created; that is, the method drives the output. For example, case study methodologies generally generate text while modeling and simulation generate highly structured, complex data formats. The research questions for this construct were:

Q1. How do researchers generate data?

Q2. What methodologies do researchers use?

To generalize the model, this survey posed two multiple choice questions that allowed for multiple answers with an open-ended "other" option to allow the respondents to provide additional information (see Table 7 below).

Table 7. Data Creation Survey Question

Question	Scale
In your research, do you use	• Data that you have **created** from observations, instruments or experiments • Data that you **gather** from other sources such as databases, vendors, or webcrawls
What research methods do you use? [check all they apply]	• Surveys • Field studies • Case studies • Direct observation in experimental situations • Analysis of instrument generated data • Analysis of existing data sets • Modeling and simulation • Text or language analysis • Other

4.1.2 Quality Control

In scientific research, data quality control is the process by which data is determined to be accurate, complete, and current (Batini & Scannapieco, 2006). This includes processes to normalize the data to allow accurate merges from disparate sources creating federated content. This process is a central concept in the e-Science Data Environment. As discussed in section 3.4, the processes used to create the federated content are often referred to in the literature as mere computation (Gray et al., 2002; Henty et al., 2008; Key Perspectives, 2010; Lord & McDonald, 2003; Lyon, 2007). The scientists interviewed in the first study disputed this characterization, but no quantitative data exists describing the amount of time and effort spent on

data quality control in research labs, which would allow one to understand the level of investment in the data. Additionally, little data exists that quantifies the types of quality control processes that are used on data. Thus, a number of unanswered research questions needed to be explored.

Q3. How much effort is expended on quality control?

Q4. What data quality control processes are used regularly?

Q5. Do researchers have data quality control criteria?

Q6. Do researchers consider data quality control to be important to their science?

A set of four survey questions was developed using multiple choice and 5-point semantic rating scales (see Table 8). All of the questions contained an open-ended "other" option to allow the respondents to provide additional information.

Table 8. Quality Control Survey Questions

Question	Scale
Which of the following processes do you run on your data?	• Data normalization (resolving scale issues, reformatting for consistency, etc.) • Data cleaning (fixing errors) • Data integration (merging data from several sources) • Instrument calibration
On average per project, how much time is spent on the data normalization, cleaning, and integration processes for a research project?	• Less than 40 hours • Between 40 and 60 hours • Between 60 and 80 hours • Between 80 and 120 hours • More than 120 hours

Do you have a formalized set of criteria for judging the quality of your data?	• Yes, almost always • Sometimes • Not generally • No, almost never • Not sure
How important is your data quality control process to your science?	• Very important • Somewhat important • Not very important • Not at all important • Not sure

4.1.3 Uniqueness

Uniqueness is an important assessment criterion for preservation. But rather than the binary assessment of uniqueness described in the literature (unique or not unique), the e-Science Data Environment describes a three-level construct that includes a level of uniqueness based on the quantity and quality of the data: that is, its level of uniformity and integration, breadth, longitudinal nature, or the added value of metadata. This uniqueness of content is an outcome of the data quality process (see section 3.3 for a fuller discussion). Testing the generalizability of this construct is important to the e-Science Data Environment.

Q7. To what extent do researchers consider their data to be unique?

A single survey question was developed using multiple choice options (see Table 9) with an open-ended "other" option to allow the respondents to provide additional information.

Table 9. Content Uniqueness Survey Question

Question	Scale
After your data is collected and any data normalization, cleaning, and integration processes for a research project are complete, please indicate which of the following statements describe the uniqueness of your data: [check all that apply]	• Observation data • Experimental data • Data is unique due to the quantity and quality of the data. • Data is unique due to the level of uniformity and integration of the data. • Data is unique due to longitudinal nature of the data. • Data is unique due to the added value of metadata. • Data is unique due to the integration of unique analysis into the data. • Data is not unique and can be recreated from the original sources. • Not sure how to describe the uniqueness of the data

4.1.4 Data Collections

As discussed in section 3.5, data collections are defined as the infrastructure, organizations, and individuals needed to provide persistent access to stored data (NSB, 2005). Although this construct of data collections, a taxonomy of the ultimate disposition of research data, is widely used in the preservation literature, the distribution of data into the three types of collections – research, community, and reference – is unknown. The definition of a data collection includes infrastructure; however, the literature using this taxonomy rarely discusses

the technologies used to support these data collections. There is little quantitative data that describes the technologies that are used for the ultimate disposition of the research data. Understanding the end-of-project disposition of data is crucial to the long-term preservation of data. The research questions for data collections were as follows:

Q8. What happens to data at the end of a project?

Q9. To what extent do repositories serve as the technology for the final disposition of data?

Q10. To what extent do researchers perceive repository data submission processes as a barrier?

Q11. To what extent was the data contribution to a repository mandatory?

Q12. To what extent are researchers able to find their data once deposited?

A set of four survey questions was developed. Rather than directly asking participants to map their data into the collection taxonomy, the survey presented a set of choices representing the types of collections. Three of the choices indicate the research collection final data disposition – deleted data, stored data within the lab, and stored data within the institution. The other two choices indicate use of community or reference collections. In addition, the participants were asked about their use of repositories. For each repository used, the participants were asked to comment on the data deposit process and to provide the reason for the deposit. To know more about the use of repositories, the participants were asked if, once they have deposited data into a repository, they were able to find it and access it again (see Table 10).

Table 10. Data Collections Survey Question

Questions	Scale
When you have completed your research, what happens to your data?	• The files are deleted when a new project needs the space. • The files are copied on to CDs or DVDs when a new project needs the space. • The files are copied to a removable hard drive when a new project needs the space. • The files are copied to a data archive within your lab or research group • The files are archived within your institution. • The files are archived in a repository specific to your scientific domain. • The files are archived in a national database. • Not sure
If you can, please name the repositories to which you have deposited data and rate how easy it was to use.	• Very easy • Easy • Neutral • Difficult • Very difficult • Not sure
Please indicate the reason that you contributed data to each repository	• Mandated by the journal in which you published • Mandated by your research institution • Mandated by your funding agency • Standard practice in your lab • Individual initiative • Not sure

If you have deposited research data into a repository, were you able to find it and gain access to it?	• I have not deposited data into a repository • I have not tried to find and access my data in a repository. • I was able to find the data and access the data easily. • I was able to find the data and access it with some amount of effort. • I was able to find the data and access it with a great deal of effort. • I was not able to find it.

4.1.5 Technical Infrastructure

Technical infrastructure – the core technologies of the researcher's institution that includes data storage, network and computing resources – is a significant component of the e-Science Data Environment (see section 3.8). The previous study provides evidence of a strong correlation between the availability of low cost, high quality, well managed storage and the type and amount of data maintained by the researchers. In the survey, quantitative measures were developed to describe more accurately the technical infrastructure and its impact on preservation. A significant component of the technical infrastructure is data management, an antecedent to preservation. To enhance the e-Science Data Environment, this study asked the following questions:

Q13. To what extent does the technology infrastructure influence the antecedents to preservation?

Q14. What threats to preservation have caused data loss?

Q15. To what extent do researchers think that they understand best practice for data management?

Q16. To what extent do researchers think that they practice best practice for data management?

Q17. To what extent are data management decisions based on funding?

Q18. Who manages data in scientific laboratories?

A set of eight survey questions was developed using either a multiple choice or a 5-point agreement scale (see Table 11). All of the questions allowed for an open-ended "other" option to allow the respondents to provide additional information.

Table 11. Data Management Survey Questions

Question	Scale
Have you lost important data due to (check all that apply)	• Lack of funding • Inadvertent human error • Malicious hacking • Mistakenly thought data was no longer needed • Equipment malfunction • Lost media • Mislabeled media • Equipment obsolescence • Software no longer recognizes data • Physical disaster (flooding, power surges, etc.) • Data corruption • I have not lost data
Do you follow standard best practice for backing up your data?	• Yes, almost always • Sometimes • Not generally • No, almost never • Not sure what is best practice for data backup

In your current research environment**, data management** is (check all that apply)	• Offered to you free of charge by your school or institution • Offered to you for a fee by your school or institution • Created and funded by your department, your lab, or your research group • Created and funded through your grants
If funding were not an issue, would you (check all that apply)	• Choose different storage technologies • Save more data • Choose different data management practices • Choose different backup strategies • Hire professional staff to manage the data
In your current research environment, **data storage** is (check all that apply)	• Offered to you free of charge by your school or institution • Offered to you at a fee by your school or institution • Created and funded by your department, your lab, or your research group • Created and funded through your grants
In your current research environment, are your computing resources (check all that apply)	• Offered to you free of charge by your school or institution • Offered to you for a fee by your school or institution • Created and funded by your department, your lab, or your research group • Created and funded through your grants
In your current research environment, is your data managed by (check all that apply)	• A professional data manager or systems administrator • Each individual who creates the data

- A graduate assistant or other student
- A combination of student help and each individual researcher

4.1.6 Context and Metadata

Context and metadata are significant components of the e-Science Data Environment. The scientists interviewed in the previous study expressed concern about their metadata –about the sufficiency of the contextual metadata, the technologies used to store metadata, their ability to find and use their data in the future, and the longevity of their metadata (see section 3.6). To generalize the e-Science Data Environment, it is important to know whether other researchers have the same concerns. Because creating metadata is considered a barrier to preservation, it is important to know the researchers' level of commitment to metadata. The survey operationalized commitment as the amount of time and money that a researcher would invest in better metadata. To explain the interaction of context with format more completely, it is necessary to understand the metadata formats that are used by researcher and how well the contextual data maps to the formats.

Q19. To what extent does metadata capture all of the contextual information that scientists have?

Q20. How do researchers perceive the sufficiency of their metadata to make data discoverable in the future?

Q21. How is the metadata stored?

Q22. Do researchers use standard formats for their metadata?

Q23. Would researchers invest time or money to improve their metadata?

A set of six survey questions was developed using multiple choice, a 5-point semantic differential scale, or a 3-point agreement scale (see Table 12). All of the questions allowed for an open-ended "other" option so the respondents could provide additional information.

Table 12. Context and Metadata Survey Questions

Question	Scale
How often do you have information about your data that is not captured in metadata?	• Almost always • Sometimes • Not generally • Almost never • Not sure
How often do you have sufficient metadata to provide all of the information needed to help you and others find your data at a later date?	• Almost always • Sometimes • Not generally • Almost never • Not sure
Is your metadata	• Stored in a database • Stored in a spreadsheet • Written in a lab notebook • Documented in a text or word processing file • Inferred from the file name and directory structure of the data files
Do you use one or more standard metadata formats? If you can, please list the standard formats you use	• Yes • No • Unsure

To help your data be more useful to you and others in the future, how much time would you be willing to spend to create more metadata for your data?	• Up to 10 minutes • Up to 20 minutes • More than 20 minutes • None
In future projects, would you consider hiring a data professional (e.g. a data librarian or data curator) to help you, your research group, or your lab create better metadata	• Yes • Perhaps • No

4.1.7 Formats

One of the most surprising findings in the e-Science Data Environment was the conflation of semantic and syntactic formats as "standard" data formats (see section 3.7). To test the generalizability of this finding, the survey asked participants to list the standard data formats that they use. There is very little information about the number of times data is converted between different formats. To understand better the impact on formats, the e-Science Data Environment model was enhanced with additional quantitative data on the frequency of data conversion.

Q24. Do researchers know what standards they use?

Q25. What is the frequency of data conversions between different formats?

A set of three survey questions was developed using multiple choice options with an open-ended "other" option to allow the respondents to provide additional information and an open-ended question for collecting the standards used (see Table 13).

Table 13. Format Survey Questions

Question	Scale
In your last research project, approximately how many times did you convert data from one format into another in your research process?	• I did not convert data at all. • Less than 3 times • Between 3 and 5 times • More than 5 times • Not sure
For your last research project, which of the scenarios below would best describe the format conversion process?	• I did not convert data at all. • I converted data from a single source into a single standard format. • I converted data from multiple sources into a single standard format. • I converted data between multiple intermediate formats before I converted into a final standard format. • I am unsure of the conversion process.
Please name the data formats that you regularly use in your research.	Open-ended question

4.1.8 Preservation Awareness

Understanding the level of awareness of data preservation among scientists is an important component in devising preservation policies, services, and educational offerings for researchers. In the previous study, the sample was too small to form any firm conclusions about the nature of the researchers' understanding of preservation issues (see section 3.9). Unanswered questions regarding researchers' level of concern, their contractual obligations to keep data

viable, and their level of commitment to preservation remained:

Q26. To what extent are researchers concerned about the longevity of their data?

Q27. To what extent are researchers concerned about preserving their data?

Q 28. To what extent are researchers committed to maintaining their data for the future?

A set of five survey questions was developed, three of which used a 5-point semantic differential scale; one used a dichotomous scale with an open-ended "other" option to allow the respondents to provide additional information; and one was an open-ended question for collecting the length of any contractual obligations to keep data usable (see Table 14).

Table 14. Preservation Awareness Survey Questions

Questions	Scale
Do you worry about the longevity of your data?	• Quite a lot • Somewhat • Not much • Not at all • Not sure
The best term to describe your level of concern about preserving your data is	• Very concerned • Moderately concerned • Slightly concerned • Not concerned at all • Not sure
Do you have any contractual obligations (through grants or other agreements) to keep your data usable for a specific length of time?	• Yes • No • Not sure
If yes, over what period of time do you need to keep your data usable?	Open ended

How important to you is making your research data available to future generations of researchers?	• Very important • Somewhat important • Not very important • Not important at all • Not sure

4.1.9 Preservation Priority Assessment

In the previous study, the scientists were uncertain about assessing preservation risks and priorities (see section 3.3). As funding agencies begin to demand data management plans and sustainability plans, the ability to assess risk and priorities becomes more important. To understand the ability of researchers to assess and prioritize their data preservation needs, the e-Science Data Environment should account for risk assessment and preservation priorities. Two questions needed to be answered:

Q29. To what extent can researchers identify data that is at risk?

Q30. To what extent can researchers identify preservation priorities?

A set of two survey questions was developed using a 5-point semantic differential scale with an open-ended "other" option to allow the respondents to provide additional information (see Table 15).

Table 15. Risk Assessment Survey Questions

Questions	**Scale**
For your research data, how easy is it for you to identify the most important data to preserve?	• Very easy • Somewhat easy • Somewhat difficult • Very difficult

	• Not sure
For your research data, how easy is it for you to identify the data that is most at risk?	• Very easy • Somewhat easy • Somewhat difficult • Very difficult • Not sure

4.2 Sample

To generalize the e-Science Data Environment model, this survey used a broad survey frame of grant awardees of the National Science Foundation (NSF). The Scholarly Database, a collation of data from many sources including journal data such as Medline, Journals of the American Physical Society, and PNAS, as well as funded grants from the National Science Foundation and the National Institutes of Health, was the source of the sample (LaRowe, Ambre, Burgoon, Ke, & Börner, 2009). This database was queried for all NSF funded Principal Investigators (PIs) from 2007 through 2010. The data base provided contact information for the official Principal Investigators; information for co-Principal Investigators was not used. From those PIs with email addresses, a set of approximately1,200 unique PIs from each of the seven NSF directorate was selected randomly. The directorates are the domain-specific divisions of NSF, each with its own funding initiatives, programs, and management. These directorates are as follows:

- Biological Sciences (molecular, cellular, and organismal biology, environmental biology)

- Computer and Information Science and Engineering (fundamental computer science, computer and networking systems, and artificial intelligence)
- Engineering (bioengineering, environmental systems, civil, and mechanical systems, chemical, and transport systems, electrical and communications systems, and design and manufacturing)
- Geosciences (geological, atmospheric and ocean sciences)
- Mathematical and Physical Sciences (mathematics, astronomy, physics, chemistry, and materials science)
- Social, Behavioral and Economic Sciences (neuroscience, management, psychology, sociology, anthropology, linguistics, and economics)
- Education and Human Resources (science, technology, engineering, and mathematics education at every level, pre-K to grey) (National Science Foundation, n.d.)

This sample of 8,400 researchers is inherently broad-based. The National Science Foundation funds research and education in most fields of science and engineering. Thus the sample of NSF awardees will come from these fields producing a large sample from many domains and disciplines ranging from theoretical to applies sciences, social sciences, and engineering. This sample is inherently biased toward high-ranking scientists – those who have been awarded one or more prestigious NSF grant(s). To allow for a greater range of researchers, the solicitation encouraged the principal investigators to forward the link to the survey to the other researchers in their labs and research communities. Because the sample was drawn from the National Science Foundation's database, it was biased toward researchers based in the U.S. Future studies could be designed to survey international researchers and U.S. researchers without NSF funding.

4.3 Demographic Categorizing Information

In the previous study, a set of *a priori* theoretically relevant categories was used to create the sample: size of laboratory, funding, scientific domain, and science type. For thus study, three of the four original theoretically relevant categories were used for demographic analysis. Because science type – big science/little science – would be difficult to explain to researchers, the study did not ask participants to attempt to determine how their domain maps to science type. In addition to the theoretical categories used in the previous study, current research institution was included. Analyzing responses by institution could provide additional insights into the technology infrastructures as discussed previously.

As with the category science type, scientific domain becomes problematic in a self-reporting survey; problems can result from granularity of domains, multiple descriptions of the same domain, and errors such as misspellings and typographical mistakes. To illustrate the issues, the term "molecular biology" was analyzed in the set of National Science Foundation principal investigators who had email addresses. The set of approximately 288,000 produced 46 different spellings, abbreviations, and combinations of molecular biology including "molec biol," "molec biology," "molecular biol," "molecular biolog," "molecular biology," "molecular biology & biochem" (see Appendix C for the full list). With potential for multiple terms to describe the same domain, the level of stratification would preclude the possibility of creating generalizable conclusions based on domain. In order to use domain as a significant categorizing element, this study used the NSF directorates as described in section 4.2 as source for the list of domains that participants can select. There is one caveat. One of the seven directorates seems to be overloaded from a theoretical perspective: Mathematical and Physical Sciences. There is reason to believe that mathematics may have a significantly different perspective on data and its

preservation needs from the physical sciences. As an example, astrophysicists have a community data repository. It is projected that the mathematical sciences will not have a high perceived need for preservation while the physical sciences will perceive preservation as important. Thus, for this study, mathematical and physical sciences will be separated into two domains: mathematical sciences and physical sciences. These eight domains will provide a reasonable stratification for the data analysis.

4.4 Validating the Survey Instrument

Venkatraman and Grant (1986) define content validity as the extent that an empirical measure reflects the content domain. In other words, content validity asks "do the questions on the survey completely cover the construct without introducing extraneous ideas?" Three heuristic evaluations and a test of the survey were conducted to ensure the validity of the constructs and the items in the survey instrument.

4.4.1 Construct Validity

For this study, content validity was tested via a panel of experts. In this first heuristic evaluation, the experts were presented with a set of construct definitions and the list of survey items. The experts were asked to identify the construct or constructs covered by the item. This methodology for construct validation is a variation on a q-sort (as described by Churchill, 1979) using a format suggested by Podsakoff (2005). Using the table matrix, the experts are able to quickly identify and code the constructs. For 20 of the 31 items, there was 100% agreement that the items measured the constructs. The remaining items had a 50% to 75% agreement, with several reviewers putting the items into multiple categories. This evaluation showed that several

of the constructs had confusing definitions. Several of the constructs were merged and redefined. A number of items were modified to address only one construct.

4.4.2 Item Construction

Care was taken when constructing the survey items. As seen in section 4.1, each item was closely tied to a both a research question and a construct. An initial set of 43 questions was refined to a set of 30 by removing redundant measures of the same dimensions of the construct. The second heuristic evaluation of the survey items by an expert survey researcher in the Indiana University Center for Survey Research[3] further refined the items, following current best practice of survey research. All questions that had been formulated as statements with which participants would either agree or disagree (please indicate your level of agreement with the following statement: I am worried about the longevity of my research data) were reworked as semantic differential scales (do you worry about the longevity of your data? with answers ranging from "quite a lot" to "not at all"). During the expert evaluation, one of the items asked about two dimensions of the same construct and was, therefore, divided to make two items increasing the number of items to 31.

The third heuristic evaluation of the survey was conducted in conjunction with the survey test. Twenty researchers, faculty, post-docs, and Ph.D. students were asked to participate in the test survey. The instructions asked these researchers to take the survey and answer three questions about their experience:

1. Were there any questions that you did not understand?
2. Were there any questions that you found confusing?

[3] http://www.indiana.edu/~csr/

3. Were there any questions that you did not want to answer? If yes, would you be willing to tell me why?

Out of the 20 researchers asked to participate, 18 took the survey and 15 returned their answers to the above questions. None of the respondents found any questions that they did not want to answer. This mitigates the minor concern from the second expert heuristic evaluation that some researchers might be embarrassed or distressed about discussing the nature and vulnerability of their data. Several respondents were confused by several questions, especially ones asking for an estimate of time or effort. The items in question were revised to be more specific so to eliminate the confusion. In addition to the answers to the questions, very helpful suggestions were offered about wording of specific items and the pagination of the web survey.

This survey was administered via the World Wide Web using the tool set used by the Indiana University Center for Survey Research, Qualtrics[4]. This tool set provides a web form for developing the questionnaire, a secured results management infrastructure, and a recruitment emailer that tracks responses. As with all software, the Qualtrics questionnaire development environment has limitations. One of those constraints forced a different item design resulting in two items rather than one. The survey instrument as instantiated in the Qualtrics software is available in Appendix D.

4.5 Data Analysis

The 43 items in this survey generated 168 numeric and character variables with an additional 34 full text elements. The data from this survey lends itself to three major types of

[4] http://www.qualtrics.com/

data analysis: general descriptive analysis, analysis of demographic subcategory significance (scientific domain, size of lab, and funding source), and textual analysis.

The descriptive statistics have been created using Microsoft Excel and the Qualtrics tool. The demographic category analysis was conducted using SPSS[5]. For the demographic categorization statistics, the data was normalized following the best practice advice provided by the Indiana University Center for Statistics Consulting: unsure and other options were removed and missing data was ignored. Cross tabulations with Chi square tests for significance, one-way analysis of variance (ANOVA) tests, and bivariate correlation analysis were used to analyze this data. In this survey, three types of qualitative questions were used – categorical, Likert or other continuous scales, and binary choice (yes or no). The types of tests performed were dependent on the type of question (see Table 16). The full list of methods used per question is available in Appendix F. For the demographic information, both the size of lab and scientific domain contain categorical data while the funding is a continuous scale.

Table 16. Analysis by Type of Question

	Category	**Likert**	**Binary**
Scientific Domain	Cross tab (Chi square)	ANOVA	ANOVA
Type of Funding	ANOVA	Correlation	Cross tab (Chi square)
Size of Lab	Cross tab (Chi square)	ANOVA	ANOVA

The findings of the data analysis are presented following standard conventions. The convention for analysis of variance (ANOVA) is to report the results of the *F* test where *F* represents the variance in the sum of squares; the subscripts denote the degrees of freedom; and *p* represents the significance $(F_{2,152} = 10.490, p < .001)$**.** The convention of cross tabulation with

[5] Version 19 (copyright 2010) from IUWare (http://iuware.iu.edu/Mac#Details/748)

Chi square (χ^2) is to report the results of the test where χ^2 represents the variance in the sum of squares, the subscript denotes the degrees of freedom, and the *p* represents the significant (χ^2_{28} = *72.867, p < .001).* Correlations will be presented with the significant (*p = .020*).

For many of the questions, an "other" option with a text box for comments was allowed. This free text data was analyzed and categorized and included in the results. When text in the "other" option matched exactly or could reasonably be matched to one of the explicit options in the survey question, no attempt was made to change the responses and include them in the explicit options. As one of the researchers in the preliminary study stated "the data is the data."

5 Results

The e-Science Data Environment survey was administered between March 27, 2011 and May 6, 2011. The potential participant pool of NSF grant awardees[6] from 2007 through 2010 was retrieved from the Scholarly Database (LaRowe et al., 2009). Of the possible 41,917 candidates, 889 did not have email addresses. The remaining 41,028 were assigned random numbers and sorted by scientific domain operationalized as the seven funding NSF directorates. The first 1,200 records of the randomized records in each directorate were included in the sample. On March 27, the initial email solicitation to the entire set of 8,400 PIs of NSF funded grants was sent. On April 3, a follow up email message was sent.

Out of the 8,400 email solicitations sent, 718 messages were not delivered due to incorrect email addresses. An additional 403 people decline to participate: 172 researchers sent email messages explaining their decisions; 283 people went to the website, read the study information sheet, and did not proceed. Thus, the original sample was reduced by 1,173, resulting in the final sample of 7,227.

The researchers who wrote to decline had three reasons why they would or could not participate: lack of time, a change in status, or lack of data. The responses ranged from apologetic ("I am sorry that I can't help you with your research – good luck!") to militant ("do not email me again, you Spammer!"). A dozen researchers had retired, been promoted to a position that precluded their participation, or joined a funding agency and felt a conflict of interest. A typical response for researchers without data is

[6] The data base provided contact information for the official Principal Investigators; information for co-Principal Investigators was not used. Thus, only official PIs were contacted.

> The nature of my research (pure mathematics) is such that I never have to deal with actual data. All the issues you mentioned (the data management practice of researchers, data quality, and the long-term retention of data) never came up in my work. Therefore I don't think I am qualified to participate in your study (personal correspondence, 2011).

From the two email invitations, 897 researchers started the survey, resulting in a response rate of 10.6% for the original sample and 12.4% for the reduced sample. Of the 897 people who started the survey, 724 completed it for a completion rate of 80.7%; the completion rate of the original sample is 8.6% and 10.0% of the reduced sample. Although this is substantially less than the 40% response projected by the Center for Survey Research, it falls well within the response rates of recent articles published in *Journal of the American Society for Information Science and Technology* that range from 3.8% for a sample of NSF researchers similar to this study (Lercher, 2010) to 10.9% for a sample of researchers from 5 universities (Niu, Hemminger, Lown, Adams, Brown, Level et al., 2010) to a very vague 5% - 10% response rate for survey of High Energy Physics researchers (Gentil-Beccot, Mele, Holtkamp, O'Connell, & Brooks, 2009).

Throughout this chapter, the results of the survey will be displayed in tables. The data will be listed in the order of the responses in the survey. For example, in Table 18 below, the domains are listed in the alphabetical order that was presented to the participants in the survey; and in Table 19, the funding options are lists in the scale order from exclusively institutionally funded through exclusively grant funded – the order of the responses in the survey. The tables each have a heading row that contains a label describing the question, the response count, and the percentage of number of respondents. For questions that allow for only one response, such as Table 18, the number of responses equals the number of respondents (the number of participants

who answered the question). For questions that allow for multiple answers, such as Tables 21 and 22, the number of responses is greater than the number of responders listed in the parenthesis in the percentage column. Using Table 21 as an example, we can see that the total number of responses is 1030. Of the 796 researchers who responded to the questions, 678 (91%) checked the first response and 315 (42%) checked the second response; thus, for questions with multiple answers, the percentages, when summed, can be greater than 100.

5.1 Sample Demographic Information

The survey was designed to gather four demographic factors: researcher role, scientific domain, funding source, and size of laboratory. Each of these factors is described below and will be used throughout the following sections.

5.1.1 Researcher Role

Of the participants in the study, 90% self-identified as principal investigators (see Table 17). The study invitation email asked the researchers to pass the message on to other researchers and students. Only 51 surveys (5.7%) were completed by people who were not in the original sample.

Table 17 – Participant Roles

Role	Responses (798)	Percentage
Principal investigator	719	90%
Researcher	29	4%
Post Doc	17	2%
Ph.D. Student/candidate	11	1%
Masters Student	2	0%
Other	20	3%

Of the 3% who self-identified as "other," 3 describe themselves as data stewards, managers, or librarians; 11 describe themselves as professors or instructors; and 2 describe themselves as data providers. The sample is significantly skewed toward PIs. Using the demographic category of participant roles for further data analysis would be unproductive, as the other roles do not have sufficient numbers to produce statistically interesting insights.

5.1.2 Scientific Domain

Scientific domain is another theoretically important category. The scientific domains of the respondents are more evenly distributed as shown in Table 18 below. However, 52% of the participants are in what have traditionally been known as the "hard science" of the physical sciences, biology, and the geosciences, while 23% are in engineering and computer science. Researchers in mathematical sciences and education are the least represented in the participants. Anecdotally (through several email responses) as well as through the comments provided by the "other" option for many of the questions, these two areas do not tend to have the volume of data as do the others.

Table 18. Participant Scientific Domain

Domain	Responses (796)	Percentage
Biological Sciences	175	22%
Computer and Information Science	84	11%
Education	33	4%
Engineering	101	13%
Geosciences	104	13%
Mathematical Sciences	60	8%
Physical Sciences	133	17%
Social, Behavioral and Economic Sciences	106	13%

5.1.3 Funding

As discussed in section 3.1.2, funding is considered to have a major effect on preservation because it can affect the amount of resources available for curation activities. For this study, funding is considered to be a continuum from exclusively funded by the researchers' institution to exclusively funded by external granting agencies with the underlying expectation that researchers with more stable funding from their institutions would have different concerns than those whose funding must be renewed regularly. The results of this survey indicate that 85% researchers are funded exclusively or primarily from grants (see Table 19). Only 8% of responders are funded exclusively or primarily from their institution.

Table 19. Participant Funding

Funding	**Responses (808)**	**Percentage**
Exclusively funded by my institution	7	<1%
Mostly funded by my institution with some grant funding	58	7%
Equally funded by grants and by my institution	35	4%
Mostly grant funded with some ongoing funding from my institution	312	39%
Exclusively grant funded	389	49%
Not sure	7	<1%

Scientific domain $(F_{7,788} = 77.207, p < .001)$ and size of lab $(F_{2,792} = 32.753, p < .001)$ are significant factors for funding source. The domains of mathematics, social science, and computer science are more likely to have institutional funding than any other group. Education, engineering, and biology as domains are the mostly likely to be funded entirely by grants. The

size of lab is also significant. The larger the group of researchers, the more likely it is to be grant funded. The individual researcher is more likely to have institutional funding.

5.1.4 Size of Lab

A second indicator of available resources for preservation activities is size of lab. As defined in section 3.1.2, this study defines a large lab as five or more researchers while a small lab has fewer than five researchers. A large majority, 82%, works in some type of lab or group setting, with 47% in a large lab and 35% in small labs; 17% of researchers work independently without a group or lab (see Table 20).

Table 20. Participant Lab Size

Lab Size	Responses (795)	Percentage
I do not work in a group or a lab	139	17%
5 or more researchers	373	47%
Less than 5 researchers	283	36%

Size of lab is significantly tied to scientific domain ($\chi^2_{14} = 201.930, p < .001$).

Biologists, computer scientists, engineers, and physical scientists tend to work in large labs while educators, mathematicians, and geoscientists work as individual researchers or in mid-sized labs.

Scientific domain, size of lab, and funding source are significant factors throughout the data and will explicitly discussed in each subsection below.

5.1.5 Research Institution

In the preliminary study used to develop the initial e-Science Data Environment model, participants were selected from three separate universities to provide a diversity of experiences.

As part of the effort to generalize the model, participants were drawn from more research institutions. The participants in this study identified affiliations with 334 unique research institutions. Of these 334 unique research institutions, 267 (80%) had a single researcher participant. Of the remaining 135 institutions, 102 institutions (31%) had between 2 and 4 participants; 26 (8%) institutions had between 4 and 9 participants; and 7 (2%) institutions had between 10 and 16 participants. The participants and their institutions are widely dispersed geographically; as well as two European countries (England and Germany), Canada, and one U.S. territory (Puerto Rico), every state in the U.S. had at least one institution represented in this study (see Figure 6). Most of the states had multiple institutions; six states had over ten participating institutions: Ohio had 12 institutions; Illinois had 13 institutions; Texas had 14 institutions; Massachusetts had 17 institutions; New York had 27 institutions; and California had 38 institutions.

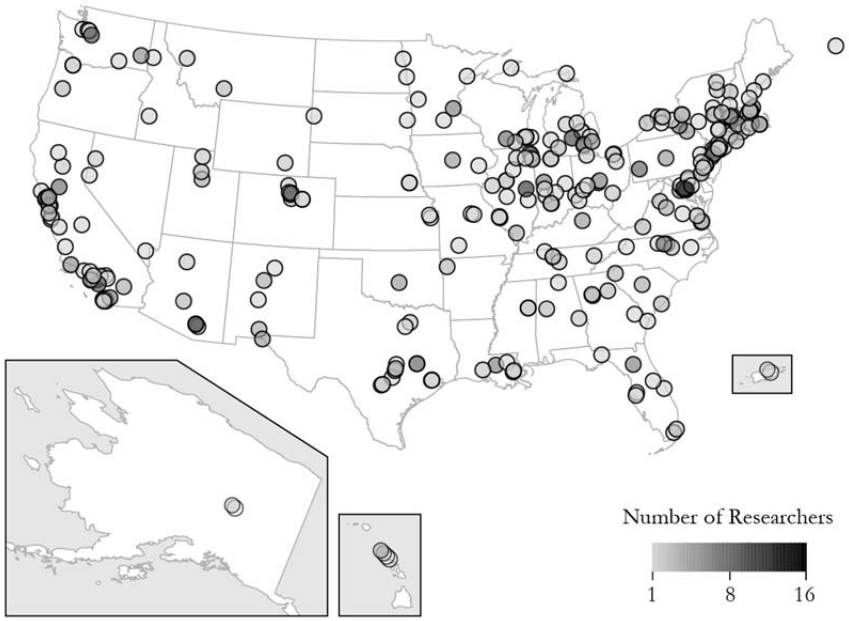


Figure 6. Geographic Distribution of Survey Participants

5.2 Data Creation

Creating data is the first step in the e-Science Data Environment model. As discussed in section 3.1, data can be *created* by observations, instruments, or experiments; or data can be *gathered* via databases, vendors, webcrawls, and other processes. Data can be created using a wide variety of research methodologies, each of which has a set of requirements, a context, and often, a specific type of output.

5.2.1 Q1. How do researchers generate data?

Of the 796 researchers who responded to this question, 91% created data and 42% gathered data (see Table 21). Scientific domain is a significant indicator for means of generating data ($\chi^2_7 = 97.928\ p < .001$). Biologists and geoscientists were more likely to use both modes to generate their research data; they created new data and gathered existing data. Physical scientists were more likely to create data; and computer scientists and mathematicians were less likely to create data. Social scientists were more likely to gather data; engineers, mathematicians, and physical scientists were less likely to gather data.

Table 21. Data Creation Methods

Methods of Creating Data	Responses (1030)	Percentage (of 796)
Data that you have created from observation, instruments, experiments, or other processes	678	91%
Data that you gathered from other sources such as databases, vendors, or webcrawls	315	42%
Other	37	5%

Researchers who are either exclusively or primarily funded via grants are more likely to create data ($F_{4,785} = 3.687, p < .000$). There is no significance between funding source and gathering data ($F_{4,785} = .184, p = .943$). Size of lab is not a significant factor in data creation methods ($\chi^2_2 = 5.420, p = .067$). Of the 5% who responded other, 34 provided additional data via the comments text box. Nine of the 34 did not use data; nine generated data via simulations; four generated data via interviews and field notes; six gathered data from previous projects or collaborators; and six generated data from published literature.

5.2.2 Q2. What methodologies do researchers use?

As discussed in section 4.1.1, data is highly dependent on the research methodologies; thus methodologies have a major impact on the e-Science Data Environment. Participants in this study were asked to identify the research methodologies that they use. Individuals could respond with multiple answers. The 782 responders used a total of 2,059 methodologies (an average of 2.7 responses per individual) and provided 98 meaningful comments (see Table 22).

Table 22. Participant Research Methodologies

Methodologies	**Responses (2076)**	**Percentage (of 782)**
Surveys	152	19%
Field studies	231	29%
Case studies	92	12%
Direct observation in experimental situations	366	46%
Analysis of instrument generated data	385	49%
Analysis of existing data sets	300	38%
Modeling and simulation	376	48%
Text or language analysis	62	8%
Other	112	14%

Research methodologies were sensitive to scientific domain (see Table 23) and the size of lab (see Table 24). Researchers who used surveys were more likely to be in education or the social sciences and to work individually. Researchers who used field studies were more likely to be in the geosciences or biology and in large or mid-sized labs. Individual researchers were less likely to use field studies as were researchers in computer science, engineering, mathematics, and physical science. Researchers who used the case study methodology were more likely to be in computer science, geosciences, social science, and education and to work individually.

Table 23. Participant Research Methodologies by Scientific Domain

Methodologies	**More Likely**	**Less Likely**
Surveys *(χ^2_7 = 119.848, p < .001)*	Education Social Science	Biology Engineering Geoscience Mathematics Physical Science
Field studies *(χ^2_7 = 172.085, p < .001)*	Biology Geoscience	Computer Science Engineering Mathematics Physical Science
Case studies *(χ^2_7 = 59.041, p < .001)*	Computer Science Education Geoscience Social Science	Biology Mathematics Physical Science
Direct observation in experimental situations *(χ^2_7 = 93.527, p < .001)*	Biology Engineering Physical Science	Education Geosciences Mathematics Social Science
Analysis of instrument generated data *(χ^2_7 = 126.361, p < .001)*	Biology Engineering Geoscience Physical Science	Computer Science Mathematics Social Science
Analysis of existing data sets *(χ^2_7 = 63.869, p < .001)*	Biology Geoscience	Education Engineering Mathematics

Modeling and simulation *(χ^2_7 = 83.468, p < .001)*	Computer Science Engineering Geoscience Physical	Biology Education Social Science
Text or language analysis *(χ^2_7 = 61.479, p < .001)*	Computer Science Education Social Science	Biology Geoscience Physical Science

Direct observation in experimental situations was more likely to be used in large labs in the biology, geology, physical science, and engineering. Individual researchers had a very low likelihood of using direct observation in experimental situations. Analysis of data generated by instruments was more likely to be used by biologists, engineers, geoscientists, and physical scientists and in large labs. Analysis of existing data sets was more likely to be used in biology and geosciences while education, mathematics and engineering were less likely to use this methodology. Computer science, engineering, geoscience, and physical science domains were more likely to use modeling and simulation as a methodology while biology, social science, and education were less likely to use it. Modeling and simulation were more likely to be used in large labs and less likely to be used by individual researchers. Text and language analysis were more likely to be used by individuals and less likely by large labs. Researchers in computer science, social science, and education were more likely to use text and language processing while researchers in biology, geosciences, and physical sciences were less likely to use this methodology.

Funding source was significant to three of the research methodologies. Researchers who were exclusively or primarily grant funded were more likely to use either direct observation in experimental situations ($F_{4,785} = 3.078, p = .014$) or analysis of instrument generated data ($F_{4,785} = 6.533, p < .001$). Researchers who were exclusively funded by their institutions were more

likely to use modeling and simulation as their methodology *($F_{4,785} = 2.682, p = .029$)*. It is possible that for this issue, funding may mirror scientific domain, as computer science researchers were more likely to be funded exclusively by their institutions and biologists were more likely to be funded exclusively by grants.

Table 24. Participant Research Methodologies by Size of Lab

Methodologies	More Likely	Less Likely
Surveys *($\chi^2_2 = 8.292, p = .016$)*	Individuals	Large Labs
Field studies *($\chi^2_2 = 7.082, p = .029$)*	Large Labs Mid-sized Labs	Individuals
Case studies *($\chi^2_2 = 6.771, p = .036$)*	Individuals	
Direct observation in experimental situations *($\chi^2_2 = 68.950, p < .001$)*	Large Labs	Individuals
Analysis of instrument generated data *($\chi^2_2 = 59.735, p < .001$)*	Large Labs	Individuals
Analysis of existing data sets *($\chi^2_7 = 2.720, p = .257$)*	n/a	n/a
Modeling and simulation *($\chi^2_2 = 31.674, p < .001$)*	Large Labs	Individuals
Text or language analysis *($\chi^2_2 = 8.360, p = .015$)*	Individuals	Mid-sized Labs

Comments from the "Other" option were analyzed and revealed four important additional categories of research methodologies including thought processes, mathematical analysis, interviews, and software development. Nearly 3% of the researchers responded using terms that described thought processes as a methodology. Typical responses were "thinking," "thinking hard," "reading and thinking," "pure thought," and "theoretical analysis." Another 1.7% used mathematical analysis methodologies that include proofs and theorem as well as mathematical

reasoning. It was anticipated that interviews would be a technique used in other methodologies such as case studies or field studies. However, 1.5% of the responders indicated that interviewing was an important methodology in and of itself. Although it was not anticipated that software development would be considered a research methodology, 1.5% of responders identified software development as their research methodology, which involves such activities as algorithm development and testing, programming computational tools, and developing data analysis tools.

As seen in Table 25, a majority of the respondents (79%) used multiple research methodologies, with 67% using between 2 and 4 different methodologies. This opens a new window into the workings of e-Science. If each of these methodologies creates multiple file formats, the complexity of the data to curate will grow with each additional methodology.

Table 25. Multiple Research Methodologies Used

Number of Methodologies	Responses (791)	Percentage
One	215	27%
Two	193	24%
Three	179	23%
Four	111	14%
Five	66	8%
Six	25	3%
Seven	2	<1%

5.3 Quality Control

In scientific research, data quality control is the process by which data is determined to be accurate, complete, and current (Batini & Scannapieco, 2006). The areas of concern for this

study were the amount of effort expended on quality control, the importance of data quality to the researchers, the processes used to insure quality, and the criteria for judging quality.

5.3.1 Q3. How much effort is expended on quality control?

The researchers in this study indicate that a substantial amount of time is spent on quality control in their scientific research. The researchers were asked to indicate the amount of time that they spent on quality control process for a recent project. As seen in Table 26, the responses fell into rough thirds: one third spent less than 40 hours; another third spent between 40 and 120 hours and the final third spent over 120 hours of effort on quality control. Both scientific domain ($\chi^2_{35} = 104.443, p < .001$) and size of lab ($\chi^2_{10} = 31.914, p < .001$) are significant, while funding source is not significant ($F_{4,724} = 8.927, p = .604$).

Table 26. Effort Expended on Quality Control for a Recent Project

QC Time	Responses (736)	Percentage
Less than 40 hours	226	31%
Between 40 and 60 hours	76	10%
Between 60 and 80 hours	77	10%
Between 80 and 120 hours	68	9%
More than 120 hours	241	33%
Other	48	7%

By domain, researchers in geoscience and biology were much more likely to have spent over 120 hours on quality control for their chosen project while researchers in physical science and mathematics were more likely to have spent less than 40 hours on quality control. Individual researchers and those in mid-sized labs were more likely to spend less than 40 hours while those

in large labs were more likely to spend more than 80 hours on quality control. Of those commenting via the "other" option, 12 indicated effort substantially higher than 120 hours, reporting times such as "6 months of 20 people," "one full time person," and "50% of project effort."

5.3.2 Q4. What data quality control processes are used regularly?

A major component of ensuring data quality is developing and applying processes to combine data from numerous sources, manipulate data to reconcile different scales of measurement, and to validate the content. The 711 researchers in this study who answered this question used an average of 2.4 processes per project including data normalization, cleaning, integration, and calibration (see Table 27).

Table 27. Quality Control Data Processes

Data Processes	Responses (1725)	Percentage (of 711)
Data normalizing (resolving scale issues, reformatting for consistency, etc.)	477	67%
Data cleaning (fixing errors)	423	59%
Data integration (merging data from several sources)	447	63%
Instrument calibration	292	41%
Other	86	12%

Quality control processes were highly sensitive to scientific domain (see Table 28). Using a cross tabulation with a chi square test for each process and scientific domain, a pattern of usage emerges. In general, mathematicians and computer scientists do not use quality control processes, as their research is not data focused, while biologists and geoscientists are more likely to use more types of process in their research. Biologists are more likely to normalize, clean, and

integrate data from multiple sources but not to calibrate instruments. Geoscientists are more likely to clean, integrate data, and to calibrate instruments. Social scientists are more likely to clean and integrate data but not to calibrate instruments. Physical scientists and engineers are more likely to calibrate instruments but not to clean or integrate data.

Table 28. Quality Control Processes by Domain Significance

QC Process	More Likely to Use	Less Likely to Use
Data normalizing	Biology	Mathematics
($\chi^2_7 = 76.440, p < .001$)		
Data cleaning	Biology	Engineering
($\chi^2_7 = 80.864, p < .001$)	Geosciences	Mathematics
	Social Sciences	Physical Science
Data integration	Biology	Computer Science
($\chi^2_7 = 79.238, p < .001$)	Education	Engineering
	Geoscience	Mathematics
	Social Science	Physical Science
Instrument calibration	Engineering	Biology
($\chi^2_2 = 147.661, p < .001$)	Geoscience	Computer Science
	Physical Science	Mathematics
		Social Science

Each of the quality control processes was sensitive to the size of lab as well (see Table 29.) As with scientific domain, a pattern of process use appears when examined by lab size. Researchers who work independently are less likely to use data quality processes than those in large labs. Funding source was significant to only one of the data quality control processes,

instrument calibration ($F_{4,785} = 3.584, p = .004$). Those researchers who are exclusively or primarily grant funded are more likely to calibrate instruments.

Table 29. Quality Control Processes by Size of Lab Significance

QC Process	More Likely to Use	Less Likely to Use
Data normalizing ($\chi^2_2 = 43.842, p < .001$)	Large Labs	Individuals
Data cleaning ($\chi^2_2 = 2.886, p = .236$)	n/a	n/a
Data integration ($\chi^2_2 = 10.708, p = .005$)	Large Labs	Individuals Mid-sized labs
Instrument calibration ($\chi^2_2 = 61.094, p < .001$)	Large Labs	Individuals

Through the comments provided in the "other" option, responders were able to identify other types of data quality control processes that are used: 21 respondents used statistical methods to determine quality; 16 respondents used specific types of analysis such as simulations, modeling, and image processing; 9 respondents used data verification such as checking field notes and running validation programs; 9 did not run quality control processes, 3 of whom said they have qualitative data and these processes are unnecessary; and 7 respondents did not have data.

5.3.3 Q5. Do researchers have data quality control criteria?

A majority of responders (61%) have some type of criteria for judging the quality of their data: 33% of responders report almost always using a set of criteria and 28% reporting using

criteria sometimes (see Table 30). Those who infrequently or never have a set of criteria have a sizable minority of 33%. Neither funding source (correlation with $p = .115$) nor size of lab ($F_{2,647} = 5.545, p = .077$) are significant to quality control criteria. Of the demographic categories, only scientific domain ($F_{7,643} = 55.916, p < .001$) was a significant factor for the quality control criteria construct. By domain, researchers in geoscience, biology, education, and physical science were more likely to regularly have quality control criteria, while researchers in social science, computer science, mathematics, and engineering were more likely to rarely have a set of criteria.

From comments in the "other" option, different views on quality criteria emerged. Several researchers indicated that the criteria were variable depending on the nature of the project, the nature of the data itself, and the specifics of the instrumentation that created the data. A pair of researcher offered differing perspectives on assumptions of data quality. One of these researchers said that he has no need for quality control because he gets his data from a well-known data repository. The other researcher states that as a data manager for a data repository, he takes all data that is deposited without regard or the ability to judge the quality.

Table 30. Criteria for Data Quality

QC Criteria	Responses (692)	Percentage
Yes, almost always	229	33%
Sometimes	193	28%
Not generally	144	21%
No, almost never	85	12%
Not sure	25	4%
Other	16	2%

5.3.4 Q6. Do researchers consider data quality control to be important to their science?

A substantial majority (88%) indicated that data quality is important to the quality of their science (see Table 31). Scientific domain was a significant factor for the importance of quality control *($F_{7,657} = 40.428, p < .001$)*: researchers in geoscience, biology, social science, education, and physical science were more likely to attribute greater importance to quality control; researchers in engineering and computer science were more likely to report that quality control was moderately important while researchers in mathematics were more likely to report that quality control was not important to their research. Neither funding source (correlation of $p = .852$) nor size of lab has significance to this question *($F_{2,661} = 1.526, p = .243$)*.

Table 31. Importance of Data Quality Control on Science

QC Importance	Responses (689)	Percentage
Very important	434	63%
Somewhat important	169	25%
Not very important	45	7%
Not at all important	17	2%
Not sure	18	3%
Other	6	<1%

5.4 Uniqueness

Uniqueness is an important assessment criterion for preservation. As discussed previously, uniqueness has been viewed throughout the literature as binary – data is unique and should be preserved; or data is derived, can be recreated, and does not be preserved (Gray et al., 2002; Henty et al., 2008; Key Perspectives, 2010; Lord & McDonald, 2003; Lyon, 2007). The results of the preliminary study show that uniqueness is more complicated than is reported in the

literature; researchers considered their quality control processes to add scientific value that creates a unique view of the data. The research question posed for this construct is a follows:

Q7. To what extent do researchers consider their data to be unique?

As did the researchers in the preliminary study, the researchers in this study rejected the simplistic binary assessment of uniqueness (unique or not unique) as described in the literature. The 747 respondents to this question confirmed the preliminary study's conclusion that uniqueness is a multidimensional construct; in fact, researchers considered their data to be unique for multiple reasons. With 1,794 responses by the 747 researchers, there was an average of 2.3 responses per researcher. In additional to the traditional definition of uniqueness (data is observational or experimental), the researchers in this study considered their data to have unique features such as integrated analysis, uniformity and quality, metadata or a longitudinal perspective (see Table 32).

All of the types of uniqueness were sensitive to scientific domain and most were sensitive to size of lab (see Appendix G for the detailed results of the cross tabulations and Chi square tests and ANOVA analysis). Biology was a significant domain for all types of uniqueness; that is, biologists were more likely to consider their data to be unique in every category. Biology was alone among the domains to be more likely to report uniqueness due to added metadata. Social scientists were more likely to report uniqueness due to the longitudinal nature of their data and the quality and quantity of their data. Geoscientists were more likely to consider their data unique due to its observational nature, its quality and quantity, its uniformity and integration, as well as the integration of analysis. Engineers and physical scientists were more likely to report their data as unique due to its experimental nature and less likely to claim uniqueness due to added metadata.

Table 32. Uniqueness of Data

Uniqueness	Responses (1725)	Percentage (of 747)
I have observation data that is unique	338	45%
I have experimental data that is unique	370	50%
Data is unique due to the quantity and quality of the data	312	42%
Data is unique due to the level of uniformity and integration of the data	132	18%
Data is unique due to the longitudinal nature of the data	136	18%
Data is unique due to the added value of metadata	118	16%
Data is not unique and can be recreated from the original sources	113	15%
Data is unique due to the integration of unique analysis into the data	117	16%
Not sure how to describe the uniqueness of this data	107	14%
Other	51	7%

Only observational data and longitudinal data were not affected by size of lab. Researchers in large labs were more likely to report uniqueness due to added metadata, integration of analysis, uniformity and integration, quality and quantity, and the experimental nature of their data. Individual researchers and those in mid-sized labs were less likely to report uniqueness due to the experimental nature of their data, the quality and quantity of their data, uniformity and integration of their data, or added metadata.

Source of funding was significant for four of the eight types of uniqueness. Researchers who are funded either exclusively or primarily though grants were more likely to report that their data is unique due to additional metadata. Researchers who were funded exclusively either from their institutions or by grants were more likely to report that they had unique experimental data or that they had data that is unique due to its quantity and quality. Those researchers who reported having observational data were more likely to be funded by grants.

Through the comments provided by the "other" option, new insights into uniqueness were brought to light. Several respondents indicated that their simulation data was more important than the original data. While in principle the simulation is possible to recreate, it is very difficult to reproduce in actuality. Others indicated that the novelty of the approach, the equipment used, the content developed, or the phenomenon under consideration was the basis for the uniqueness of their data.

5.5 Data Collections

The construct of data collections describes a taxonomy of the ultimate disposition of research data: research collections, community collections, and reference collections (NSB, 2005). Research data collections refer to the output of a single researcher or lab during the course of a specific research project. Community data collections generally serve domain or other well defined area of research. At the highest level, reference data collections are broadly scoped, widely disseminated, well funded collections that support the research needs of many communities (NSB, 2005). Although the definition of a data collection includes infrastructure, the literature using this taxonomy rarely discusses the technologies used to support these data

collections. There is little quantitative data that describes the technologies that are used for the ultimate disposition of the research data.

5.5.1 Q8. What happens to data at the end of a project?

As discussed in section 3.5, none of the researchers in the preliminary study would have used the National Science Board's data collection taxonomy; so rather than directly asking participants to map their data into this taxonomy, the survey presented a set of choices representing the types of collections: five of the options indicated the research collection; one of the options described a community collection; and one of the options described a reference collection (see Table 33). The 716 researchers who responded to this question used multiple strategies and technologies to maintain their data at the end of a project. A very small percentage of researchers (3%) deleted data at the end of a project, and a large percentage of researchers (72%) copied their data to removable media such as CDs, DVDs, and hard drives. A small majority (51%) of the researchers had some type of technology supplied by their institution for data storage, while 27% of the researchers used community or reference collections for the long-term storage of their data. Through the comments provided through the "other" option, 91 researchers provided additional information. Several researchers who deleted data at the end of a project wanted to qualify their choice by stating that they only deleted a small number of files – primarily intermediate results files that they thought were inconsequential. Almost 6% of the researchers indicated that they do nothing to their data once a project is complete; the data remains as it was. Less than 1% of the respondents indicated that they published their data either in a journal paper, as supplemental files with their journal paper, or on a personal or department website.

Table 33. End of Project Disposition

End of Project Disposition	Collection Type	Responses (1725)	Percentage (of 716)
Files are deleted	N/A	23	3%
Files are copied on to CDs or DVDs	Research	242	34%
Files are copied to a removable hard drive	Research	269	38%
Files are copied to a lab data archive	Research	247	35%
Files are archived within your institution	Research	111	16%
Files are archived in a domain repository	Community	96	13%
Files are archived in a national database	Reference	98	14%
Not sure	N/A	23	3%
Other	N/A	91	13%

A large majority of the researchers, 81%, indicated that they used one or more method of storing of their research collections: 49% used one of the methods; 24% used two of the methods; 6% used three; 1% used four. Unexpectedly, 5.7% did not choose any research collection options but did choose one or more of the other collections. That is, these researchers have contributed to community or reference collections but did not choose any of the individual research collection options.

Scientific domain was a significant factor in end of project disposition of research data (see Table 34). Out of the seven options available to researchers in this survey, the only one that was not sensitive to domain was file deletion. For their research collections, biologists were more likely to use removable media as well as lab or department data archives. Computer scientists were less likely to use any removable media but more likely to use an institutional archive. Geoscientists and physical scientists were more likely to use removable hard drives; geoscientists

were also more likely to use institutional data archives. Social scientists and educators were more likely to use department data archives for their research collections.

For community collections, both biologists and geoscientists were more likely to report contributing to domain data collections while physical scientists, social scientists, and educators were less likely to report using community collections. For reference collections, again, both biologists and geoscientists were more likely to report contributing to national data collections while computer scientists, physical scientists, social scientists, and educators were less likely to contribute. Mathematicians and engineers were less likely to report using any of the options listed for their data.

Table 34. End of Project Disposition by Scientific Domain

End of Project Disposition	Collection Type	More Likely	Less Likely
Files are deleted *(χ^2_7 = 10.587, p = .158)*	N/A	n/a	n/a
Files are copied on to CDs or DVDs *(χ^2_7 = 30.065, p < .001)*	Research	Biology	Computer Science Engineering Geoscience Mathematics
Files are copied to a removable hard drive *(χ^2_7 = 37.772, p < .001)*	Research	Biology Geoscience Physical Science	Computer Science Mathematics
Files are copied to a lab archive *(χ^2_7 = 25.164, p = .001)*	Research	Biology Computer Science Education Social Science	Engineering Geoscience Mathematics
Files are archived within your institution. *(χ^2_7 = 15.850, p = .027)*	Research	Computer Science Geoscience	Engineering Mathematics Social Science
Files are archived in a domain repository	Community	Biology Geoscience	Education Engineering

(χ^2_7 = 67.292, p < .001)			Mathematics Physical Science Social Science
Files are archived in a national database. *(χ^2_7 = 58.634, p < .001)*	Reference	Biology Geoscience	Computer Science Education Engineering Mathematics Physical Science Social Science

Size of lab was significant to three of the seven disposition options. CDs or DVDs *(χ^2_2 = 10.992, p = .004)*, removable hard drives *(χ^2_2 = 14.133, p = .001)*, and department data archives *(χ^2_2 = 29.216, p < .001)* were all sensitive to lab size. In all cases, researchers in large labs were more likely to use these methods for end of project data disposition while researchers in mid-sized labs or individual researchers were less likely to report using these methods.

Source of funding was significant to reference collections *($F_{4,785}$ = 1.379, p = .013*). Those researchers who were exclusively grant funded were more likely to contribute data to a national database. In all other categories, there was no statistical significance by funding source.

5.5.2 Q9. To what extent do repositories serve as the technology for the final disposition of data?

Of the 716 researchers who provided information about the final disposition of their data as described above, 154 (21.5%) had archived their data in a community or a reference collection, thus implying that approximately 78% have not deposited data into a formal data collection. If institutional archives are included, the number of researchers using repositories increased to 234 (32%) and reduced the implied non-repository users to 69%. As seen in Table 41 in section 5.5.5 below, 347 out of 580 respondents (59%) have not deposited data into a

repository. Although the exact percentage varied depending on the question asked and the number of respondents, it is reasonable to conclude that formal repositories such as community and reference collections are not the primary technologies for the final disposition of research data.

When asked to name the repositories that they had used, 240 researchers provided 388 responses. As these responses were reviewed, it became clear that the word "repository" has multiple interpretations and meanings. In addition to listing the anticipated well known reference data collections as well as a variety of lesser known, but easily identifiable community data collections, the researchers identified a number of other types of repositories. During the analysis, six primary categories of repositories emerged: commercial, community, institutional, journal, personal, and reference (see Table 35).

Table 35. Types of Repository

Type of Repository	Responses (388)	Percentage
Commercial	12	3%
Community	74	19%
Institutional	42	11%
Journal	25	6%
Personal	55	14%
Reference	142	37%
No repositories used	2	1%
Unable to Categorize	15	4%
Blank	21	5%

Repositories were categorized as institutional if they were supported by an institution and had no clear community affiliation. Included in this category were such repositories as library systems, institutional repository software systems (DSpace), and a university controlled resource ("[my university's] storage service"). Institutional repositories were identified in 42 of the responses (11%). Repositories were categorized as commercial if they were a service offered by a for-profit corporate entity. Examples of commercial repositories are Amazon's Web Services (AWS)[7], Google Documents[8], and Dropbox[9]. Only 3% of the repositories used could be classified as commercial. Repositories were categorized as community or reference if they fit the National Science Board (2005) definitions. Community collection repositories account for 19% of repository usage, and reference collection repositories account for 37%. Journal publishers have increasingly been requiring authors to submit data with accepted papers prior to publication (Kowalczyk & Shankar, 2010). In this study, 6% of repositories used by respondents could be classified as journal publishers and included both commercial publishers such as Elsevier and non-commercial publishers such as PubMed.

Of the 240 researchers, 39 (16%) apparently interpreted "repository" to mean any storage device and listed compact disks, hard drives, flash drives, or other removable storage media as the repository to which they had deposited data. These personal data collections on removable media will not be considered as repositories in this paper or included in further analysis. An additional 13 researchers (5%) provided 21 responses where the repository name was left blank while indicating the ease of use and motivation data for these unnamed repositories. These

[7] http://aws.amazon.com/

[8] https://docs.google.com/

[9] http://www.dropbox.com/

responses will be considered in the generalized assessment of ease of use and motivation but will not be used in types of repository analysis. Another 15% of responses had insufficient information for categorization; as examples, the text "metadata" and "culture collections" were entered into the repository field but cannot be categorized or considered as a repository. These, too, will be excluded from further analysis.

When the analysis of this data was complete, the final list contained 221 separate repositories used by 199 researchers. This set of 221 categorized repositories will be used throughout the remainder of this section. The diversity of these repositories was unexpected. Of the 221 repositories named, 183 (83%) were referenced only once. Of the remaining 14% of repositories that were identified multiple times, only 8 had more than five responses. Within this list of most frequently named repositories are four reference collection repositories, three community collection repositories, and one journal repository (see Table 36).

Table 36. Most Frequently Named Repositories Used By Respondents

Repository	Type	Responses	Percentage
NCBI – The National Center for Biotechnology Information	Reference	59	26%
LTER – Long Term Ecological Research Network	Community	8	4%
IRIS – Incorporated Research Institutions for Seismology	Reference	6	3%
UNAVCO – The University NAVSTAR Consortium	Community	5	2%
PDB – The Worldwide Protein Data Bank	Reference	5	2%
NGDC – The National Geophysical Data Center	Reference	5	2%
ICPSR – The Interuniversity Consortium for	Community	5	2%

Political and Social Research			
Ecological Archives	Journal	5	2%

5.5.3 Q 10. To what extent do the researchers perceive repository data submission processes as a barrier?

For all repositories, a majority (66%) of researchers found it either easy or very easy to contribute data to a repository while 14% found it difficult or very difficult (see Table 37). Neither size of lab *($F_{4,375} = 1.212, p = .057$)* nor scientific domain *($F_{7,228} = 5.995, p = .622$)* is significant for repository ease of use.

Table 37. Overall Repository Ease of Use

Rating	**Responses (380)**	**Percentage**
Very Easy	120	32%
Easy	128	34%
Neutral	78	21%
Difficult	46	12%
Very Difficult	8	2%
Total	380	

When analyzed by type of repository, a more nuanced view of ease of use emerged (see Table 38). Commercial repositories and journal repositories had a much higher ease of use rating than do reference and community repositories. The most obvious explanation for this variance is the amount of effort (i.e. metadata) required during the submission process. The commercial repositories required only an active account and a named file. Beyond the published

paper, minimal metadata is required for journal repositories such as PubMed (National Center for Biotechnology Information, 2008) and Ecological Archives (Ecological Society of America, 2011). Both community and reference repositories demanded a substantial amount of metadata that could certainly impact the perceived ease of use.

Table 38. Ease of Use by Repository Type

	Commercial		Community		Institutional		Journal		Reference	
	#	%	#	%	#	%	#	%	#	%
Very Easy	5	42%	19	26%	13	31%	12	48%	32	23%
Easy	3	25%	23	32%	16	38%	8	32%	50	36%
Neutral	2	17%	18	25%	10	24%	2	8%	35	25%
Difficult	2	17%	11	15%	2	5%	2	8%	21	15%
Very Difficult	0	0%	2	3%	1	2%	1	4%	2	1%
Totals	12		73		42		25		140	

5.5.4 Q11. To what extent was the data contribution to a repository mandatory?

For all repositories, two of the top three motivations to contribute to a repository were mandates – either from a funding agency or from the researcher's peers (see Table 39). These motivations are sensitive to scientific domain ($\chi^2_{35} = 92.919, p < .001$) and to size of lab ($\chi^2_{10} = 23.471, p = .009$). Biologists were more likely to have a journal mandate data deposit and were less likely to have a mandate from their funding agency or have a personal motivation.

Computer scientists were more likely to have a personal motivation and less likely to have mandates from their journals or funding agencies. Geoscientists were more likely to have a mandate from their funding agencies and less likely to have a mandate from their journals. Physical scientists were more likely to report contributing to a repository due to standard lab or department practice and less likely to have mandates from journals or funding agencies. Social scientists were more likely to have a personal motivation. Motivation to contribute to repositories was not sensitive to source of funding *($F_{5,226} = 7.647$, $p = .073$).*

Table 39. Motivations to Contribute to Repositories

Motivations	**Responses (375)**	**Percentage**
Mandated by journal	59	16%
Mandated by research institution	13	3%
Mandated by funding agency	93	25%
Standard practice in research lab/group	92	25%
Individual initiative	106	28%
Other	12	3%

When analyzed by type of repositories, the motivations became clearer (see Table 40). For commercial repositories, the primary motivation was personal initiative; that is, the researchers who seek out and use these commercial services were interested in finding a safe place for their data. The primary motivation to deposit data with a journal repository was, unsurprisingly, a mandate from the journal. For reference and community repositories, the main motivation was a funding agency mandate. It was anticipated that funding agency mandates would be a motivating force. The finding that standard practice within a research lab or group

would have such a large impact on repository usage was unexpected. For community, intuitional, journal, and reference repositories, lab practice contributed to more than 20% of the motivations.

Table 40. Motivation for Repository Use by Type

	Commercial		Community		Institutional		Journal		Reference	
	#	%	#	%	#	%	#	%	#	%
Mandated by journal	2	17%	5	7%	0	0%	12	48%	37	26%
Mandated by research institution	0	0%	1	1%	6	14%	0	0%	0	0%
Mandated by funding agency	2	17%	25	34%	5	12%	3	12%	51	36%
Standard practice in research lab/group	1	8%	16	22%	11	26%	6	24%	29	21%
Individual initiative	5%	50%	24	32%	19	45%	4	16%	22	16%
Other	1	1%	3	4%	1	2%	0	0%	1	1%
Totals	12		74		42		25		140	

5.5.5 Q12. To what extent are researchers able to find their data once deposited?

Because preservation and access are tightly coupled, providing access to archived data is an important function of repositories (Lesk, 2008). The researchers in this study were asked about their experiences retrieving their data from a repository once it had been deposited. Of the

590 researchers who responded to this question, 28% were able to find their data relatively easily (see Table 41). The ability to access data once deposited is sensitive to scientific domain ($\chi^2_{28} = 85.843, p < .001$). Biologists and geoscientists are more likely to have reported easy access to data and less likely to report that they have not deposited data. Computer scientists, engineers, mathematicians, social scientists, and educators were more likely to report that they had not deposited data into a repository. Neither size of lab ($\chi^2_8 = 5.189, p = .737$) nor source of funding ($F_{4,569} = .511, p = .954$) is significant to the ability to access data after deposit.

Table 41. Ability to Access Data After Deposit in Repository

Access after deposit	Response (590)	Percentage
I have not deposited data into a repository	347	59%
I have not tried to find and access my data in a repository.	34	6%
I was able to find the data and access the data easily.	162	28%
I was able to find the data and access it with some amount of effort.	38	7%
I was able to find the data and access it with a great deal of effort.	6	<1%
I was not able to find it.	3	<1%

5.6 Technical Infrastructure

Technical infrastructure is defined as the core technologies of the researcher's environment including data storage, network, and computing resources. Data management, an antecedent to preservation, is defined as the process by which data in which is stored, maintained, administered, and protected. As such, data management is considered part of the technical

infrastructure. Technical infrastructure is a multilevel construct and can occur at the institution, the department, the lab, or the individual level. As developed in the preliminary study, there was evidence of a correlation between the availability of low cost, high quality, well managed storage and the type and amount of data maintained by the researchers.

5.6.1 Q13. To what extent does the technology infrastructure influence the antecedent to preservation?

The researchers in this study were asked to identify the sources of their technical infrastructure. The sources reflected the levels of the multidimensional construct – institution, work unit (such as department or lab), or individual. The elements of the technical infrastructure include data storage, data management, and computing environments. Three questions, one for each of the infrastructure elements, were posed; each allowed multiple responses per question. The results, combined in Table 42, show a complex combination of sources per element. Data storage and computing environments have substantially more institutional support, both as a free service and as a fee-for-service, than does data management. Fee-for-service offerings from institutions were not as common as was anticipated. Departmental support for all three elements was greater than anticipated and was consistent across all three elements. Grants funded all three elements of the technology infrastructure for half of the researchers in this study.

Additional insight into the technical infrastructure came from the comments via the "other" option. The researchers in this study used a number of creative options to fund their technology infrastructure needs. Institutionally supplied discretionary research funds were used by some to either buy storage media and/or computers as well as to hire students for data management support. Nearly 2% of researchers reported using their own funds to purchase technology infrastructure. Others (2%) used external repositories for a combination of storage

and data management. Still others cited collaboration partners as providers of technology infrastructure. Several researchers report no supported from their institutions at all ("Nothing is provided at [my institution]") or that services are in the process of be developed ("[my institution] is just now instituting data storage capacity because NSF requires it"). One resourceful researcher stated that he uses his "…own personal laptop for most of my work, but [I] also use computers in my wife's lab at another university."

Table 42. Technical Environment Components

	Data Storage		**Data Management**		**Computing Environment**	
	Responses (1021)	**% (of 726)**	**Responses (887)**	**% (of 721)**	**Responses (1119)**	**% (of 728)**
Offered to you free of charge	299	41%	110	15%	338	46%
Offered to you for a fee	70	10%	34	5%	91	13%
Created and funded by your department or lab	209	29%	187	26%	260	36%
Created and funded through your grants	354	49%	373	52%	392	54%
Not sure	35	5%	109	15%	12	2%
Other	54	7%	74	10%	26	4%

The technical environment was sensitive to lab size, scientific domain, and funding source (see Appendix H for tables). Individual researchers or those who worked in mid-sized labs were more likely to use free, institutionally supported data storage and computing

environments. Researchers in larger labs were more likely to create their own infrastructure including data storage, data management, and computing environments either through their department/lab or their grant funding. Biologists and physical scientists were more likely to develop all three elements their infrastructure via their grant funding. Geoscientists and physical scientists also had a higher likelihood of department/lab support for data storage and computing environments. Social scientists and mathematicians were more likely to use free institutionally provided computing environments. Social scientists also were more likely to use free institutionally provided data storage. Computer scientists were less likely to use institutionally provided support or to fund their infrastructure via their grants but were more likely to have department support for data management.

Researchers who had mixed funding, that is who were financed by a combination of institutional and grant support, were more likely to have both their data storage and data management provided by their work units such as departments or labs and their computing environment offered for no cost by their institutions. Researchers who were either exclusively or primarily supported by their institutions were more likely to report that data storage was provided at no cost by their institutions. Researchers who were either exclusively or primarily grant funded were more likely to provide their own technical infrastructure through their grants.

5.6.2 Q14. What threats to preservation have caused data loss?

The literature is rich with categorizations of the threats to preservation. Many of these threats, such as human error, obsolescence, lost media, and physical disaster, are human failures and can be considered as failures of data management. Other threats such as software, hardware, or media faults can be described as technology failures. The researchers in this study were asked to identify the causes of any known data loss based on the list of threats developed in section

2.5.2 (see Table 43). Scientific domain had significance to only two of the threats while size of lab was significant to six of the twelve threats. Of the 716 respondents to the question on data loss, 50% reported some type of data loss. Human based data management failures were prominent in the list of reported reasons for data loss. Direct loss through inadvertent human error was the largest cause of data loss (30%). This threat was sensitive to the size of lab ($\chi^2_2 = 18.329, p < .001$). Researchers in large labs were more likely to report data loss due to inadvertent human error while individual researchers and those who work in mid-sized labs were less likely to have suffered such a loss.

Data loss due to obsolescence is another failure of data management. As described previously, because technology changes rapidly, obsolescence is an inherent property of technology and must be managed. When data is lost due to either software or hardware obsolescence, it is a failure of management. For the researchers in this study, 23% suffered data loss due to obsolescence – 11% with hardware obsolescence and 12% with software obsolescence. Obsolescence was sensitive to size of lab: software ($\chi^2_2 = 7.211, p = .027$) and equipment ($\chi^2_2 = 9.005, p = .011$) obsolescence were both more likely to be reported by large labs than individual researchers or by those in mid-sized labs. Two scientific domains, geoscience and physical science, were sensitive to equipment ($\chi^2_7 = 14.545, p = .042$) obsolescence.

Lack of disaster preparedness (Barateiro et al., 2008; Rosenthal et al., 2005; Rosenthal et al., 2004) does not seem to be a major threat to the researchers in this study, as only 4% reported loss of data due to physical disasters. Two reasonable conclusions may be drawn from this result. It is possible that the research centers and data centers in major universities have implemented business resumption plans that have mitigated this threat. It is also possible that this finding is

the result of happenstance, that the participants in this study simply have not experienced a large-scale disaster.

Although malicious attack is widely accepted as a threat to preservation (Altman et al., 2009; Baker et al., 2005; Barateiro et al., 2008; Rosenthal et al., 2005; Rosenthal et al., 2004) and is perceived as a significant and imminent threat in the literature, only 1% of the researchers in this study reported any loss due to malicious hacking. There are at least two possible explanations for the low incidence of malicious attack: one is the low reward for hacking scientific data (as opposed to credit card companies or other financial institutions), and the second is the improved security of both servers and networks at research institutions.

As described in section 2.3, incidents of data loss due to misplaced media have been well publicized; 10% of the researchers in this survey reported to have lost media and the data thereon. In addition, two researchers reported via the "other" option that they lost data because equipment was stolen. The loss of the physical media is not discussed as a threat in any of the preservation literature found to date. In their work on preservation threats, Rosenthal and colleagues (2005) discussed the rapid rate at which removable media becomes obsolete but do not address the issue of physical management of the media itself. Baker and colleagues (2005) contended that the major issue with removable media is the separation of the media from the reader, e.g. having a pile of floppy disks without a computer with an internal disk reader. Although not in the preservation literature, the threat of data loss due to lost media is a concern for businesses (Lor & Snyman, 2005). Removable media management needs to be included in the discussion of threats to preservation.

Loss of context is a major threat to preservation (Baker et al., 2005). Data sets with minimal or no metadata are virtually lost (Rumsey, 2010; Swan & Brown, 2008). The researchers in this

study confirm this finding. Mislabeled media constitutes a loss of context and metadata. In this study, it accounted for 5% of the data loss scenarios. Seven of the commenters via the "other" option reported that the lack of metadata caused a loss of access to data. All of these researchers equated the loss of access via metadata to the loss of data.

Of the 716 respondents, 91 (13%) experienced data loss due to corruption. Data loss due to corruption is difficult to categorize because it may have multiple causes, described by Baker and colleagues as multiple, cascading, and compounding failures (2005). Data corruption can be caused by media failure, software malfunctions, and human error such as partial overwrites and poor file management. Size of lab was a significant factor in data corruption *($\chi^2_2 = 19.003, p < .001$)*: researchers in large labs were more likely to report data loss due to corruption while individual researchers were less likely.

Equipment malfunction, the only technology threat covered in this study, was the second most common cause of data loss reported in this study; 24% of the researchers reported data loss from equipment failures. Equipment malfunction was sensitive to scientific domain *($\chi^2_7 = 26.945, p < .001$)* and to lab size *($\chi^2_2 = 25.914, p < .001$).* Geoscientists and physical scientists were more likely to report data loss by equipment failure while computer scientists, mathematicians and educators were less likely to have lost data due to equipment failure. Researchers in large labs were more likely to report data loss due to equipment failure while individual researchers and those in mid-sized labs were less likely.

Lack of funding was reported by 5% of the respondents. Neither scientific domain nor size of lab was significant to this issue; however, funding source *($F_{4,785} = .395, p = .048$)* was; those researchers who were equally funded by institution and grants were more likely to report data loss by lack of funding.

Of the 716 respondents to the question on data loss, 50% reported no data loss. Individual researchers and those in mid-sized labs were more likely to report no data loss while researchers in large labs were less likely ($\chi_2^2 = 10.689, p = .005$). Although this could be a positive indication of data management in action, it cannot be taken as proof of good practice. As one of the commenters in the "other" option stated, "I don't *think* that I have lost data!"

Table 43. Data Loss

Data Loss	Responses (1242)	Percentage (of 716 responders)
Lack of funding	36	5%
Inadvertent human error	216	30%
Malicious hacking	6	1%
Mistakenly thought data not needed	49	7%
Equipment malfunction	173	24%
Lost media	73	10%
Mislabeled media	34	5%
Equipment obsolescence	76	11%
Software no longer recognizes data	88	12%
Physical disaster	29	4%
Data corruption	91	13%
I have not lost data	355	49%
Other	21	3%

5.6.3 Q15. To what extent do researchers think that they understand best practice for data management?

Because data management is complex and is closely tied to the type of data, the long term needs of the researchers, and the technologies available, it is difficult to codify and reduce best practice to a survey question. However, data management is well known to be part of the entire

data lifecycle (Green & Gutmann, 2007; Loshin, 2009). As such, the item used to understand the extent to which researchers know data management best practice measures the understanding of data management timing; that is, the question asked researchers to identify the point in the research data life cycle at which data management becomes important. A majority of researchers (55%) reported that data management is important throughout the life cycle, beginning with the creation of the data (see Table 44). An additional 17% indicated that data management should begin at the onset of data analysis, a relatively early stage in the data lifecycle. By this measure, a large majority, 72%, had a reasonable understanding of the importance of managing data early in the lifecycle. Of the 720 respondents, 151 (21%) did not convey a reasonable understanding of data management as they indicated that data management is important only as the data life cycle is concluding: 4% as the analysis is complete; 8% as research papers are written; 3% as data is archived; 4% as data is missing; and 2% are unsure.

Table 44. Data Management within the Research Lifecycle

Data Management Timing	Responses (720)	Percentage
Managing data is not important to my research.	50	7%
When the data is created.	398	55%
When the analysis begins.	121	17%
When the analysis is complete.	29	4%
When papers are being written.	56	8%
When the data needs to be archived.	23	3%
When I need to find something and can't remember where it is.	26	4%
Not sure	17	2%

Size of lab affected the perception of the need for data management within the data lifecycle ($\chi^2_{12} = 25.388, p = .013$): individual researchers were more likely to consider data management to be unimportant while those in mid-sized labs considered data management to be important when the data analysis began. Responses from researchers in large labs did not vary from the expected distribution. Scientific domain also had a significant impact this issue ($\chi^2_{42} = 130.438, p < .001$). Both computer scientists and mathematicians were more likely to find data management was unimportant to their research. Engineers were more likely to find data management was important when writing their papers. Biologists, social scientists, and educators were more likely to find data management important at data creation. Funding source was not significant ($F_{4, 693} = 5.850, p = .580$).

5.6.4 *Q16. To what extent do researchers think that they practice best practice for data management?*

A significant component of data management, and indeed data preservation, is the assurance of multiple copies of data in disparate locations. The primary method of creating multiple copies is creating backups. When asked to identify if they followed standard best practice for backing up their data, 55% of the researchers reported that they almost always did. An additional 28% of the researchers indicated that they did sometimes follow best practice. Only 8% reported that they rarely or never followed best practice for data backup. An alarming 10% said they did not know best practice (see Table 45).

Scientific domain was significant to the perceived use of best practice ($\chi^2_{28} = 42.922, p = .011$). Computer scientists and geoscientists were significantly more likely to report that they always used best practice. Engineers were more likely to report that they sometimes used best

practices and were less likely to report always using best practice. Neither size of lab (χ^2_8 = *10.883, p = .208)* nor source of funding *($F_{4, 697}$ = 3.993, p = .634*) was a significant factor.

Table 45. Use of Best Practice for Backup

Best Practice	Responses (725)	Percentage
Yes, almost always	375	52%
Sometimes	204	28%
Not generally	42	6%
No, almost never	15	2%
Not sure what is best practice	72	10%
Other	17	2%

5.6.5 Q17. To what extent are data management decisions based on funding?

The researchers were asked to identify data management practices that they would change if money were not an issue. Of the 583 responders, 55 (9%) reported through the "other" option that they would not make any changes; a typical response was as follows: "Our current system is adequate; all problems so far have been due to human error." If funding were not an issue, the remaining 91% the researchers would make different choices about the amount of data they stored, the technologies used to store that data, and the processes and staff by which they managed their data (see Table 46). The most common change researchers would institute is to hire professional staff to manage their data.

Table 46. Data Management Funding Options

Funding Options	Responses (999)	Percentage (of 583)
Choose different storage technologies	144	25%

Save more data	144	25%
Choose different data management practices	167	29%
Choose different backup strategies	181	31%
Hire professional staff to manage the data	276	47%
Other	87	15%

Only two of the options were sensitive to scientific domain: choosing different data management practices *($\chi^2_7 = 16.524, p = .021$)* and hiring professional staff to manage the data *($\chi^2_7 = 29.788, p < .001$).* Biologists and educators were more likely to want to change both options if funding were available, while social scientists would change only data management practices and geoscientist would hire data professionals. Mathematicians and physical scientists were less likely to want to change their data management practices or to hire a data professional. Computer scientists were less likely to want to hire data professionals. Only one of the options was sensitive to the source of funding *($F_{4, 785} = 3.093, p = .008$)*; researchers who were exclusively or primarily funded by grants were more likely to want to hire a data professional.

As seen in Table 47, all but one of the options were sensitive to size of lab. While saving more data does not vary based on size of lab, the desire to hire data professionals as well as to choose different storage technologies, different data management practices, and different backup strategies were more likely among researchers in large labs. Individual researchers and those in mid-sized labs were less likely to want to implement changes.

Table 47. Data Management Funding Options By Size of Lab

Funding Options	More Likely	Less Likely
Choose different storage technologies (χ^2_2 = 13.460, p = .001)	Large Labs	Individual Mid-size Labs
Save more data (χ^2_2 = 4.891, p = .087)	n/a	n/a
Choose different data management practices (χ^2_2 = 7.266, p = .026)	Large Labs	Individual Mid-size Labs
Choose different backup strategies (χ^2_2 = 14.167, p = .001)	Large Labs	Individual Mid-size Labs
Hire professional staff to manage the data (χ^2_2 = 6.726, p = .035)	Large Labs	Individual Mid-size Labs

5.6.6 Q18. Who manages data in scientific laboratories?

From this study, it was clear that individual researchers were primarily responsible for managing their own data; of the 721 respondents, 408 (57%) reported that they had sole responsibility for the data (see Table 48). An additional 38% researchers reported that they had student support for their data management functions. Dedicated data management staff, either professional staff or student, was reported by 25% of the researchers.

Table 48. Data Management Staffing Models

Data Management Staffing	Responses (961)	Percentage (of 721)
A dedicated professional data manager or systems administrator	108	15%
Each individual who creates the data	408	57%
A dedicated graduate assistant or other student	72	10%

A combination of student help and each individual researcher	272	38%
Not sure	21	3%
Other	80	11%

In the comments provided through the "other" options, several other models emerged. Four researchers indicated that they had collaborations with other institutions; one of the researchers explained the process: "This is a collaborative effort in my group. I provide the data management environments, oversee and help when necessary, but other researchers are active participants." The other three researchers gave their data to a collaborating partner institution who then took on all responsibility for data management. Ten of the researchers reported that they had technical staff who manage the data, some of whom are paid directly from grants while others are paid by the department. The largest number of comments came from those researchers who were personally responsible for data management. The researchers who commented about data management responsibilities expressed more emotion than in any other response. While a number of exclamation marks were used in responses to other questions, this is the only issue which elicited an emoticon and had negative tones; when explaining who is responsible for data management, this researcher stated, "I do it. :-(." Another stated, "I get stuck with this."

Scientific domain was a significant factor in data management staffing (see Table 49). Biologists were more likely either to have access to a professional data manager or to manage their own data but were less likely to have a graduate student as a data manager. As with biologists, geoscientists were more likely to have access to a professional or to manage their own data; but unlike biologists, they were also more likely to have graduate students to assist with the data management tasks. Physical scientists were more likely to have access to a professional

data manger while engineers, mathematicians, and social scientists were less likely. Engineers were more likely to manage their own data without the help of graduate students, while computer scientists and social scientists were less likely to manage their own data and use graduate students.

Table 49. Data Management Staffing by Scientific Domain

	More Likely	**Less Likely**
A professional data manager or systems administrator *(χ^2_7 = 23.883, p = .001)*	Biology Geosciences Physical Science	Engineering Mathematics Social Science
Each individual who creates the data *(χ^2_7 = 33.645, p < .001)*	Biology Engineering Geosciences	Computer Science Mathematics Social Sciences Education
A graduate assistant or other student *(χ^2_7 = 14.184, p = .048)*	Computer Science	Biology
A combination of student help and each individual researcher *(χ^2_7 = 17.751 p = .013)*	Geosciences Social Science	Engineering Mathematics

Size of lab was also a significant factor in data management staffing (see Table 50). In general, researchers in large labs had more options for managing their data than did individual researchers. Researchers in large labs were more likely than individuals to have access to a professional data manager, to manage their own data, and to have some student support for data management. Those researchers in mid-sized labs were less likely to either have profession support or to manage their own data. Funding source was not significant *($F_{4,785}$ = .187, p = .173).*

Table 50. Data Management Staffing by Size of Lab

	More Likely	Less Likely
A professional data manager or systems administrator *(χ_2^2 = 16.283, p < .001)*	Large Labs	Individual Mid-sized Labs
Each individual who creates the data *(χ_2^2 = 19.202 p < .001)*	Large Labs	Individual Mid-sized Labs
A graduate assistant or other student *(χ_2^2 = 3.212, p = .201)*	n/a	n/a
A combination of student help and each individual researcher *(χ_2^2 = 10.578, p = .005)*	Large Labs	Individual

5.7 Context and Metadata

Context and metadata are often considered to be synonymous in the literature; but in this research, context is considered to be more inclusive than metadata. Context describes the relationships of the data content to its environment (CCSDS, 2002). Metadata is codified information about data, generally using one or more predetermined structured representational formats. Metadata is the method by which data is described so that it can be understood within its context. The preliminary study raised questions about the sufficiency of the contextual metadata, the technologies used to store metadata, researchers' ability to find and use their data in the future, the longevity of their metadata, the durability of their metadata formats, and the researchers' commitment to creating good metadata (see section 3.6).

5.7.1 Q19. To what extent does metadata capture all of the contextual information that scientists have?

In the preliminary study, a number of the directors of scientific labs reported that they often had more contextual data about their data than they could express within their metadata schemes. To verify this finding, the researchers in this study were asked about the frequency of this phenomenon (see Table 51). Of the 692 researchers who responded, 384 (55%) reported that they have more data than they can represent in their metadata scheme: 29% almost always do, and 26% sometimes do. A surprisingly large percentage (21%) was unsure. In general, these findings confirm the preliminary study; researchers frequently had more contextual data than can be accounted for in their metadata schemes. None of the demographic categories were significant: scientific domain *($F_{7,521} = 8.213$, $p = .396$)*, size of lab *($F_{2,525} = 1.453$, $p = .524$)*, or funding source *($p = .405$)*.

Table 51. Information About Data Not Captured in Metadata

Information not in Metadata	Response (692)	Percentage
Almost always	204	29%
Sometimes	180	26%
Not generally	64	9%
Almost never	81	12%
Not sure	144	21%
Other	19	3%

5.7.2 Q20. How do researchers perceive the sufficiency of their metadata to make data discoverable in the future?

As discussed earlier, perceived usefulness and accessibility to the data is an important component of preservation (Lesk, 2008). It is, therefore, important to consider the sufficiency of metadata for access. As seen in Table 52, when asked if they had sufficient metadata to provide all of the information required for discovery over time, 60% of the responding researchers expressed that they did, either always (33%) or sometimes (27%). Nearly equal numbers were unsure (18%) or pessimistic (20%) about the sufficiency of their metadata.

Table 52. Sufficient Metadata for Reuse

Sufficient Metadata	Responses (689)	Percentage
Almost always	224	33%
Sometimes	189	27%
Not generally	66	10%
Almost never	66	10%
Not sure	126	18%
Other	18	3%

Scientific domain was a significant factor for this issue: *($F_{7,537} = 32.551, p < .001$).* Biologists, physical scientists, geoscientists, and social scientists were more likely to indicate confidence that they had sufficient metadata while computer scientists, educators, mathematicians, and engineers were less likely to have confidence in their metadata. There was a significant correlation between funding source and metadata sufficiency (*$p = .012$*); researchers who were funded by their institutions were less likely to report that they had sufficient metadata

while those who were funded by grants were more likely to report sufficient metadata. Size of lab was not a significant factor *($F_{2,541} = .141$, $p = .933$).*

5.7.3 Q21. How is the metadata stored?

In order for metadata to be useful in preservation, it must be both explicit and actionable; that is, it must be specifically expressed in a format that can be used by computers for automatic processes. In the preliminary study, the research lab directors reported that a substantial amount of their metadata resided in implicit or non-actionable media such as lab notebooks, file names and directory structures, or text files. For this study, researchers were asked to describe how their metadata was stored (see Table 53).

Table 53. Metadata Storage

	Responses (1398)	**Percentage (of 681)**
Stored in a database	195	29%
Stored in a spreadsheet	226	33%
Written in your lab notebook	229	34%
Documented in a text or word processing file	325	48%
Inferred by the file name and directory structure	241	35%
Other	62	9%
Not sure	120	18%

Of the 681 researchers who responded, 421 (62%) used an explicit and actionable technology for storing their metadata: 29% in a database and 33% in a spreadsheet. In addition, 82% of the researchers had explicit but non-actionable metadata stored in lab notebooks or text

or word processing files. Text and word processing files are considered to be non-actionable file formats since there is not an explicit structure to capture individual data elements for extraction.

A substantial number of researchers, 35%, had implicit metadata that is inferred by file names and directory structures. This implicit metadata can, with effort, be extracted programmatically under certain circumstances. For example, digital libraries have used implicit metadata in file names and directory structures for years to create structural metadata for scanned books and other page-oriented materials (Stanford University Libraries, 2005). Files would have a compound name with a common prefix identifying a volume number and a suffix with the page sequence number (i.e.,12345_001 as the first page image of volume 12345). Consistency of the structure must be maintained for successful metadata extraction. Although the implicit structural data may be extracted, creating metadata that describes formats, experimental conditions, samples, and other intellectual content from file names and directory structures would be exceedingly difficult.

Scientific domain was a significant indicator for metadata storage (see Table 54). Biologists and geoscientists presented a mixed message about metadata. They were more likely to store their data in explicit, actionable formats such as databases and spreadsheets as well as to use the traditional lab notebook that is unactionable. Geoscientists were more likely to use text files to store their metadata. Physical scientists were more likely to use formats that make automatic processing more difficult by using either lab notebooks (explicit and unactionable) or file names and directory structures (implicit and perhaps actionable). Surprisingly, computer scientists were less likely to use explicit actionable metadata storage technologies than other researchers; they were no more likely to use databases than any other researchers but were much less likely to use spreadsheets. They were much more likely to use implicit file organizations and

unactionable text files. And of course, mathematicians were unlikely to have metadata and thus had little use for metadata storage technologies.

Table 54. Metadata Storage By Scientific Domain

	More Likely	**Less Likely**
Stored in a database *(χ^2_7 = 54.968, p < .001)*	Biology Geosciences	Education Engineering Mathematics
Stored in a spreadsheet *(χ^2_7 = 72.213, p < .001)*	Biology Geosciences	Computer Science Mathematics
Written in your lab notebook *(χ^2_7 = 79.400 p < .001)*	Biology Geosciences Physical Sciences	Computer Science Education Mathematics Social Sciences
Documented in a text or word processing file *(χ^2_7 = 25.083, p = .001)*	Computer Science Geosciences Social Sciences	Biology Engineering Mathematics
Inferred by the file name and directory structure *(χ^2_7 = 30.467, p < .001)*	Computer Science Physical Science	Biology Mathematics

Size of lab was also significant to the storage of metadata (see Table 55). In all cases, individual researchers were less likely to store their metadata in any of the methods proposed in the survey while researchers in large labs were more likely to do so. Researchers in mid-sized labs varied from the expected distribution in only three cases: they were less likely to store their metadata in a database and were more likely to use both lab notebooks and file names and directory structures.

Two of the metadata storage options were sensitive to funding source: spreadsheets *($F_{4, 785} = 1.941, p = .048$)* and lab notebooks *($F_{4, 785} = 1.953, p = .050$)*. Spreadsheets were more likely to be used by researchers who were primarily funded by grants. Lab notebooks were more likely to be used by researchers who were either exclusively or primarily grant funded. It is possible that the results of funding source are mirroring the results of the scientific domain. Biology, the most likely to use these two storage methods, is also more likely to be grant funded.

Table 55. Metadata Storage by Size of Lab

	More Likely	**Less Likely**
Stored in a database *($\chi^2_2 = 12.284, p = .002$)*	Large Lab	Individual Mid-sized Lab
Stored in a spreadsheet *($\chi^2_2 = 10.403, p = .006$)*	Large Lab	Individual
Written in your lab notebook *($\chi^2_2 = 26.988\ p < .001$)*	Large Lab Mid-sized Lab	Individual
Documented in a text or word processing file *($\chi^2_2 = 11.836, p = .003$)*	Large Lab	Individual
Inferred by the file name and directory structure *($\chi^2_2 = 7.301, p = .026$)*	Large Lab Mid-sized Lab	Individual

An unanticipated result was the large number of researchers who reported uncertainty about their metadata (18%). The "other" option provided some additional details about metadata storage: 2% of researchers responding had no need for metadata; nearly 2% of the researchers used self documenting stand file formats that include metadata (FITS, EML, etc.); 1% used

published peer reviewed journal papers as their primary metadata storage; five researchers used wiki or web sites for metadata storage; and two used email messages.

5.7.4 Q22. Do researchers use standard formats for their metadata?

When asked if they used standard metadata formats, a substantial majority (57%) reported that they did not and 29% reported that they were unsure. Only 11% indicated that they did use standard metadata formats. Neither funding source *($F_{4,662} = 3.094, p = .141$)* nor size of lab *($\chi^2_2 = 1.542, p = .462$)* was a significant factor in for use of standard metadata formats. Scientific domain was a significant factor *($\chi^2_7 = 16.023, p = .025$)* with biology and geosciences more likely to use standard metadata formats and engineering and social sciences less likely to use standard metadata formats.

Table 56. Use of Standard Metadata Formats

Standard Formats Usage	Response (671)	Percentage
Yes	75	11%
No	385	57%
Not sure	195	29%
Other	16	2%

Of the 75 researchers who reported using standard metadata formats, 73 provided additional textual information describing the metadata standard used. This data was analyzed to determine the types of standard formats used; that is, to categorize the formats as proprietary formats, as syntactic data formats, or as domain-specific semantic and syntactic formats. Twelve of the responses were too generic to encode, consisting of such statements as “multiple standards

depending on source" or "we are shifting to electronic notebooks." The remaining 61 responses were categorized by the type of format (standard, proprietary, and syntactic). Of the 63 responders, 21 (33%) use proprietary formats for generic functions such as spreadsheets and for specific scientific applications; 33 researchers (52%) reported using domain specific standards; five researchers (8%) used syntactic standard for text or tabular data; and two researchers (3%) reported using the formats specified by their data repositories.

5.7.5 Q23. Would researchers invest time or money to improve their metadata?

To gauge the importance of and commitment to metadata, the researchers in this study were asked if they would be willing to spend time to make their metadata more useful and if they would be willing to hire professional staff to create metadata. In general, the researchers indicated a great willingness to contribute time to enhance the quality of their metadata (see Table 57); nearly 80% would spend some amount of time: 20% would spend up to 10 minutes; 18% would spend up to 20 minutes; and 40% would spend more than 20 minutes enhancing their metadata.

The amount of time that researchers were willing to spend to improve their metadata was sensitive to both scientific domain ($\chi^2_{21} = 58.688, p < .001$) and size of lab ($\chi^2_6 = 12.960, p = .044$); however, it was not sensitive to funding source ($F_{4,618} = 2.853, p = .674$). Biologists, social scientists, and educators were more likely to be willing to spend more than 20 minutes; engineers and physical scientists were more likely to be willing to spend up to ten minutes, while mathematicians were more likely to be unwilling to spend any time on metadata. Individual researchers were skewed to either end of the question; that is, they were either more likely to spend more than 20 minutes on metadata or no time on metadata. Researchers in large labs were more likely to spend up to 10 minutes on metadata.

The researchers who provided comments through the "other" option indicated that they had a very strong commitment to creating good metadata but that the amount of time spent was dependent on the type of project, the amount of data, the types of data and the number of different types of data. One respondent gave a very insightful comment: "Everything takes huge amounts of time – more like 20 hours than 20 minutes. It would be nice to have funding to do this."

Table 57. Willingness to Improving Metadata

Time	Responses (682)	Percentage
Up to 10 minutes	139	20%
Up to 20 minutes	125	18%
More than 20 minutes	275	40%
None	87	13%
Other	56	8%

Although the researchers in this study were willing to spend time to enhance their metadata, they were not as willing to spend any of their research funds to hire a data professional (such as a data librarian or data curator) to help create better metadata; 20% would be willing, 32% might be willing, and 44% are not willing (see Table 58). Both scientific domain *(χ^2_{14} = 50.490, p < .001)* and size of lab *(χ^2_4 = 10.116, p = .039)* were significant indicators for this question while funding source was not significant *($F_{4,683}$ = 1.839, p = .627*). Two of the scientific domains had a more positive view towards hiring a data professional: biology and education researchers were more likely to answer yes. Physical science researchers were more likely to answer no. Individual researchers were more likely to say no. Researchers in large labs

were more likely to say yes or perhaps, while mid-sized labs were more likely to say perhaps or no.

Table 58. Willing to Hire Metadata Professional in the Future

	Responses (693)	Percentage
Yes	140	20%
Perhaps	219	32%
No	305	44%
Other	29	4%

The comments from the "other" option provided some insight. Many of the comments indicated that the researchers would not have enough work for a full time person. Others did not think that a non-scientist could understand the nature of the data or would have problems with the large number of very different types of formats required in their research. Still others doubted that any of their funders would be willing to support data professionals.

5.8 Formats

Digital data is represented in a file format, which is defined as the internal structure and encoding that facilitate computational processing as well as rendering for human use (Brown, 2006). Three questions regarding the formats of the researchers' data were asked: what data formats are used; what is the frequency of data conversion; and what scenario best describes the conversion process.

5.8.1 Q24. Do researchers know what standards they use?

The participants of this study were asked to list the file formats that were used in their research; the question allowed free text responses, so respondents could list multiple formats. The 390 researchers who responded to this question provided 921 usable answers. These responses were analyzed to determine the type of format; that is, to categorize the formats as proprietary formats, as syntactic data formats, as standard formats, or as domain-specific semantic and syntactic formats. As well as categorizing the type of format, the data was analyzed to determine the function of the format such as for instrument data, statistical data, spreadsheets, databases, time sequences, biological sequencing, and others. With some syntactic data formats such as text (.txt), ASCII, or binary, it was not possible to determine the function as the formats were too generic to provide context for the function.

Of the 921 responses, 12% were very general and did not name a specific format but a description or a research function; when possible, these general answers were kept. As examples, general responses such as "various image formats," "numerous proprietary instrument data," and "spreadsheets and tables" were encoded as generic types with a specific function. But some responses were too general to be useful and are not included in the 912 useful responses; for example, "digital," "measurements," and "alphanumeric" did not provide sufficient information to allow for encoding; as well, the acronyms used in three responses could not be resolved with certainty and were discarded.

It was expected that most of the formats used by principal investigators of National Science Foundation funded projects would be domain specific syntactic and semantic standards. However, only 12% of the formats reported were these domain specific standards including such formats as MIRIAD (Multichannel Image Reconstruction Image Analysis and Display for

telescope data), FITS (astronomical data), NetCDF (array data), SEG-Y (seismic data), NBRF (genome sequences), and FASTA (nucleotide or peptide sequences). The largest proportion, 48% of reported formats, were proprietary; that is these formats are closed; opaque; require specific licensed software to open, read, or write; or have patent or other ownership restrictions. The most common of the proprietary formats are from Microsoft Office products such as Excel, Word, and Access. Statistical package internal formats such as SPSS, SAS, and R were heavily used. Desktop database packages FoxPro and FileMaker Pro were also reported. The second largest type of format reported by the participants in this study is syntactic standards; 22% of the formats cited defined the structure of the file but included no semantic meaning of the data. The most common of these syntactic standards are comma separated values (.csv), tab delimited, ASCII, plain text (.txt), and binary. Traditional standard file formats such as TIFF, JPG, PDF, and MEG in all of its varieties for audio and video constituted only 9% of the formats reported. Just over 1% of researchers in this study used paper as their data format.

In addition, these formats were analyzed to determine their primary functions when possible. The major functions that these formats support are spreadsheets (17%), statistical processing (13%), digital images (9%), text (8%), database functions (6%), and geographical processing (4%).

5.8.2 Q25. What is the frequency of data conversions between different formats?

With the large number of formats that researchers reported using, it was not surprising that 73% of the 732 respondents converted data at least once in a recent research process with 40% converting fewer than three times and 33% converting three or more times (see Table 59). Frequency of format conversion was sensitive to both scientific domain ($\chi^2_{21} = 65.150, p < .001$) and the size of the researchers' lab ($\chi^2_6 = 19.010, p = .004$) although it was not sensitive to

funding source ($F_{4,670} = 4.472, p = .348$). Biologists and geoscientists were more likely to convert data formats more than 5 times while engineers and mathematicians do not convert at all. Computer scientists and physical scientists were more likely to convert formats fewer than 3 times. Social scientists were more likely to convert data between 3 and 5 times. Individual researchers were more likely not to convert between formats; researchers in mid-sized labs were more likely to convert fewer than 3 times; and researchers in large labs were more likely to convert more than 5 times.

Table 59. Format Conversions

Number of conversions	Responses (732)	Percentage
I do not convert data at all	147	20%
Fewer than 3 times	291	40%
Between 3 and 5 times	122	17%
More than 5 times	119	16%
Not sure	42	6%
Other	11	2%

The complexity of the potential specific scenarios for converting data is captured in Table 60. Of the 717 respondents to this question, 26% either did not convert data or were unsure of the process. The use of different conversion scenarios was nearly evenly distributed; the simple scenarios – a single source to one or more standard formats – have the same percentage of use as the more complex scenarios of multiple sources. The conversion scenarios were sensitive to funding source ($F_{4,626} = 49.730, p = .016$). Researchers funded by grants were more likely to have multiple conversions. Size of lab was not a significant factor ($\chi^2_8 = 12.161, p = .144$) for

conversion scenarios; however, these scenarios were sensitive to scientific domain (χ^2_{28} = *72.867, p < .001).* Engineers and mathematicians were more likely to have no format conversions. Computer scientists and physical scientists were more likely to convert from a single source into a single standard format. Biologists were more likely to convert data between multiple intermediate formats before converting into a final standard format. Geologists and social scientists were more likely to use two conversion scenarios – one that converts from multiple sources into a single format as well as to convert into multiple intermediate formats.

Table 60. Conversion Scenarios

Conversion Scenarios	**Responses (717)**	**Percentage**
I do not convert data at all	148	21%
I convert data from a single source into a single standard format	126	18%
I convert from a single source into multiple standard formats	117	16%
I convert data from multiple sources into a singe standard format	132	18%
I convert data between multiple intermediate formats before I convert into a final standard format	113	16%
I am not sure of the conversion process	39	5%
Other	42	6%

Researchers' comments in the "other" option indicated that nearly 4% convert multiple sources to multiple intermediate formats to multiple final formats. One researcher's comment

confirmed the complexity: “All of the [options listed in the question] are true even within one project!”

5.9 Preservation Awareness

In the preliminary study, the directors of the research labs surveyed expressed deep concern about the longevity of their data, both the actual data and the contextual metadata. However, the sample was too small to form any firm conclusions about the nature of the researchers’ understanding of preservation issues. This study sought to quantify researchers’ level of concern, their contractual obligations to keep data viable, and their level of commitment to preservation.

5.9.1 Q26. To what extent are researchers concerned about the longevity of their data?

The 769 responders were nearly evenly divided between those who were concerned about the longevity of their data (47%) and those who indicate lower levels of concern (48%) (see Table 61). The majority of the respondents (64%) had moderate or slight concerns.

Concern for the longevity of data varied significantly between scientific domains *($F_{7,723} = 49.951, p < .001$).* Mathematicians and educators were much less concerned than the other researchers. Size of lab was also a significant factor *($F_{2,728} = 19.551, p < .001$).* Researchers in large labs were significantly more concerned with the longevity of their data than those in mid-sized labs. Individual researchers were the least concerned about longevity. Funding source and the concern for the longevity of data were significantly correlated ($p = .008$). Researchers who were exclusively or primarily grant funded were more likely to have greater longevity concerns than researchers who were exclusively or primarily institutionally funded.

The comments from the “other” option indicated that this question may not have solicited

the information originally intended. Several researchers stated that they were not concerned about data longevity because they had developed an archival plan: "No - I have arranged for it to be archived in a museum"; "[a data archive] has assumed responsibility for preserving the data, so I'm not worried:" [my data] will be archived;" "Not perticularly [sic] as we have a complex system to safeguard and preserve data!" Thus, being unconcerned does not necessarily mean a lack of awareness but may indicate a concern that led to action.

Table 61. Concern about Longevity

Longevity Concern	Responses (769)	Percentage
Quite a lot	108	14%
Somewhat	256	33%
Not much	239	31%
Not at all	129	17%
Not sure	14	2%
Other	23	3%

5.9.2 Q27. To what extent are researchers concerned about preserving their data?

Unlike concern for longevity of data, the respondents' preservation concerns were weighted toward low concern – 45% expressed high or moderate concern while 61% expressed slight or no concern – but the pattern of a large bubble in the center that had some level of concern was consistent (see Table 62).

The scientific domain *($F_{7,732} = 59.018, p < .001$),* size of lab *($F_{2,737} = 25.376, p < .001$),* and funding source *($p = .049$)* categorization for preservation concerns were significant and were consistent with the concerns for longevity as described above. Mathematics and education were

much less concerned with preservation than the other domains. Researchers in large labs were significantly more concerned with preservation than those in mid-sized labs or individual researchers. Researchers who were exclusively or primarily grant funded were more likely to have greater preservation concerns than researchers who were exclusively or primarily institutionally funded. As with concern for longevity, the comments from the "other" option were illuminating. One respondent stated, "I was concerned enough to do something about it."

Table 62. Preservation Concerns

Preservation Concern	Responses (768)	Percentage
Very concerned	105	14%
Moderately concerned	236	31%
Slightly concerned	237	31%
Not concerned at all	163	21%
Not sure	10	1%
Other	17	2%

5.9.3 Q 28. To what extent are researchers committed to maintaining their data for the future?

Concern for the longevity of data and the preservation of data are rather abstract concepts that may not seem important to individual researchers. Measuring the impact of the consequences of preservation could provide a more realistic understanding of the attitudes toward preservation. Although Moore (2008) characterized digital preservation as a method of communication with both the future and the past, it has not been known if researchers feel an obligation to communicate with the future, that is, to make their data available to future generations. A large majority of researchers in this study did feel an obligation to the future. As

seen in Table 63, 80% thought that it is very or somewhat important to provide access to their research data for researchers yet to come, indicating a commitment to some type of preservation. Scientific domain *($F_{7,730} = 35.136, p < .001$),* size of lab *($F_{2,735} = 11.307, p < .001$),* and funding source *($p = .007$)* were significant factors for this obligation to the future. By scientific domain, geoscientists, biologists, and physical scientists were more likely to attribute high importance to making their data available to the future. Social scientists, engineers, and computer scientists were more likely to place a moderate level of importance to making their data available to the future; and mathematicians and educators were the least likely to place importance on preserving data for the future. Individual researchers were less likely to think that preservation for the future was important. Researchers who were exclusively and primarily grant funded were more likely to be concerned about making their data available to the future, while those who were funded exclusively or primarily by their institutions were less concerned.

Table 63. Importance to make data available to future

Importance For Future	Responses (764)	Percentage
Very important	321	42%
Somewhat important	294	38%
Not very important	93	12%
Not important at all	31	4%
Not sure	11	1%
Other	14	2%

In the preliminary study, several scientists had contractual obligations to provide access to their data for a specific period of time and were concerned about their ability to keep their data

viable for the duration of that contract. Only 15% of the respondents to the second survey had such contractual obligations (see Table 64). Funding source *($F_{4,582} = 2.325, p = .005$)*, size of lab *($\chi^2_2 = 11.468, p = .003$)*, and scientific domain *($\chi^2_7 = 26.958, p < .001$)* were significant to this issue. Researchers who were funded either exclusively or primarily by grants were more likely to have contractual obligations to provide access to the future. Researchers in large labs were significantly more likely to report a contractual obligations; and individual researchers were significantly less likely to report a contractual obligation. Biologists were more likely to report contractual obligations while mathematicians and engineers were the least likely.

It is interesting to note that 19% of the responding researchers were unsure of any external requirements on their data. It was anticipated that the high number of unsure responses would correlate to researchers who were not the principal investigators; however, the percentages of the unsure respondents were virtually the same as the total sample – 90% of the unsure responses came from principal investigators.

Table 64. Contractual Obligations

Contractual Obligation?	Responses	Percentage
Yes	115	15%
No	477	62%
Not sure	147	19%
Other	28	4%
Total	767	100%

The comments from the "other" option provided additional information that extended the understanding of contractual obligations by describing some of the condition of these contracts.

Several researchers described the requirement to keep data "until last manuscript published;" others described the condition that data must be made available as soon as the quality control process was complete; still others described a cooperative arrangement with external organizations to preserve the original data while the researchers are required to maintain access to their analysis, models, and simulations. A small number of researchers had additional conditions that included the requirement to embargo the data or to destroy the data after a specific time.

Of the 767 researchers in the study who responded to the question about contractual obligations, 142 (18.5%) provided specifics on the length of these contracts or general comments about contractual obligations. This data was entered as free text and has been analyzed and coded into meaningful categories (see Table 65). For those responding, 36% had obligations to keep their data between 1 and 5 years. Another 14% were obligated to keep their data between 10 and 20 years, while 23% were obligated to keep their data in perpetuity. An additional 4% were required to deposit their data in a specific data repository. Quite a few researchers, 16%, had an obligation to keep their data but did not provide a specific length of time. A couple of responses indicated an obligation for a specific contract such as the "duration of cooperative agreement." But most of the comments reflected confusion and uncertainty about their obligations: "we agreed to make it accessible, no specific timeline;" "[the] policy is unclear;" "time not stipulated."

Table 65. Length of Contractual Obligations

Length of Time	Responses (142)	Percentage
1 - 2 years	5	4%
3 Years	21	15%
5 Years	24	17%
6 - 10 years	12	8%
10 - 20 years	8	6%
In Perpetuity	32	23%
Not Specified	23	16%
Specific Archive Required	5	4%
Public Access Required	5	4%
Unsure	4	3%
No Data	7	5%

5.10 Preservation Priority Assessment

Within the digital preservation literature, assessing preservation priorities is a significant topic. Much has been written on preservation priorities, focusing on assessment criteria; this dissertation has contributed to this discussion as seen in sections 2.2 and 3.3. Understanding the process by which researchers determined their priorities has had less focus. In the preliminary study, the scientists were uncertain about assessing preservation risks and priorities; determining the relative importance of their numerous files and understanding the vulnerability of the different formats is an important component of preservation awareness. This study looked at

researchers' understanding of risk and priority setting.

5.10.1 Q29. To what extent can researchers identify data that is at risk?

Of the 765 researchers who responded to this question, 508 (66%) answered that they could, with some level of ease, identify the data that is at risk, while 20% thought that they would have difficulty determining which data is at risk (see Table 66).

Although size of lab was not a significant factor for risk assessment *($F_{7,652} = 1.676$, $p = .296$),* scientific domain was significant for this question *($F_{7,648} = 18.419$, $p < .001$).* Researchers in mathematics and geosciences were more likely to find risk identification very easy, while researchers in social science, biology, physical science, and computer science were more likely to find this somewhat easy. Researchers in engineering and education found this to be a difficult task.

Funding source was correlated to risk assessment ($p = .020$). Researchers who were exclusively or primarily funded by their institutions were more likely to report that it would be easy or very easy to identify the data most in need of preservation, while researchers who were exclusively or primarily grant funded were more likely to report that this was a more difficult task.

Although a majority of the researchers thought this would be an easy task, it is not clear that they had the necessary understanding of the risks to make this assessment, as indicated by the findings of this study on the paucity of preservation quality data formats used (see section 5.8.1). Further study, testing the ability of researchers to judge correctly the preservation risks of their data, could be useful.

5.10.2 Q30. To what extent can researchers identify preservation priorities?

A large majority of the researchers (73%) thought that they would not have difficulty in

determining their most important data. Scientific domain was significant for this question *($F_{7,688} = 17.558, p < .001$).* Researchers in mathematics, geoscience, and the social sciences were more likely to find prioritizing data easy, while researchers in biology, physical science, and computer science were more likely to find this task to be moderately easy. Researchers in engineering and education were more likely to find this task difficult. Neither size of lab *($F_{7,692} = 2.926, p = .116$)* nor funding source (*$p = .083$*) was a significant indicator for identifying preservation priorities.

The comments provided through the "other" option explained the general risk assessment parameters for these researchers. Several of the researchers aligned their preservation priorities with their publications; data that is cited in peer reviewed journals is important. Other indicated that all of their data is important and should be preserved.

Table 66. Preservation Priority Assessment

	Ability to Identify At-Risk Data (765)		**Ability to Identify Important Data (765)**	
Very Easy	224	29%	249	33%
Somewhat easy	284	37%	303	40%
Somewhat difficult	122	16%	116	15%
Very difficult	27	4%	29	4%
Not sure	78	10%	39	5%
Other	30	4%	29	4%

Although the researchers indicated that they could more easily identify their most important data (73%) than the data that is most at risk (66%), the majority clearly has confidence

in their ability to prioritize their preservation needs. This finding conflicts with the preliminary study. It is possible that the difference in confidence in preservation assessment could be attributed to the difference in the positions of the respondents. The current study queried principal investigators while the preliminary study queried lab and program directors. A reasonable explanation could be that researchers would have a better understanding of their own data than their directors and that the directors would have a larger perspective, thinking about all of the data within their purview.

5.11 Moving from Results to a New Model of Preservation

The data from this survey has been analyzed and has provided a number of interesting results. Metadata standards were rarely used by researchers outside of biology and geology. Few researchers have deposited data into community or reference data collections. For those who have used these data collections and their repository infrastructure, the barriers to input data were lower than anticipated. The multi-dimensionality of uniqueness that was developed in the preliminary study has been upheld and expanded. In the chapter that follows, these insights will be more fully explored and will be used to expand, extend, and generalize the e-Science Data Environment that was introduced in section 3.2.

6 Discussion

The Digital Data Research Environments Study produced results that require a reevaluation of the e-Science Data Environments model developed in Chapter 3 (see Figure 7 below). With the data provided from the new study, the model can be enhanced to more fully describe the interactions between the components, the antecedents to preservation, and the barriers to preservation.

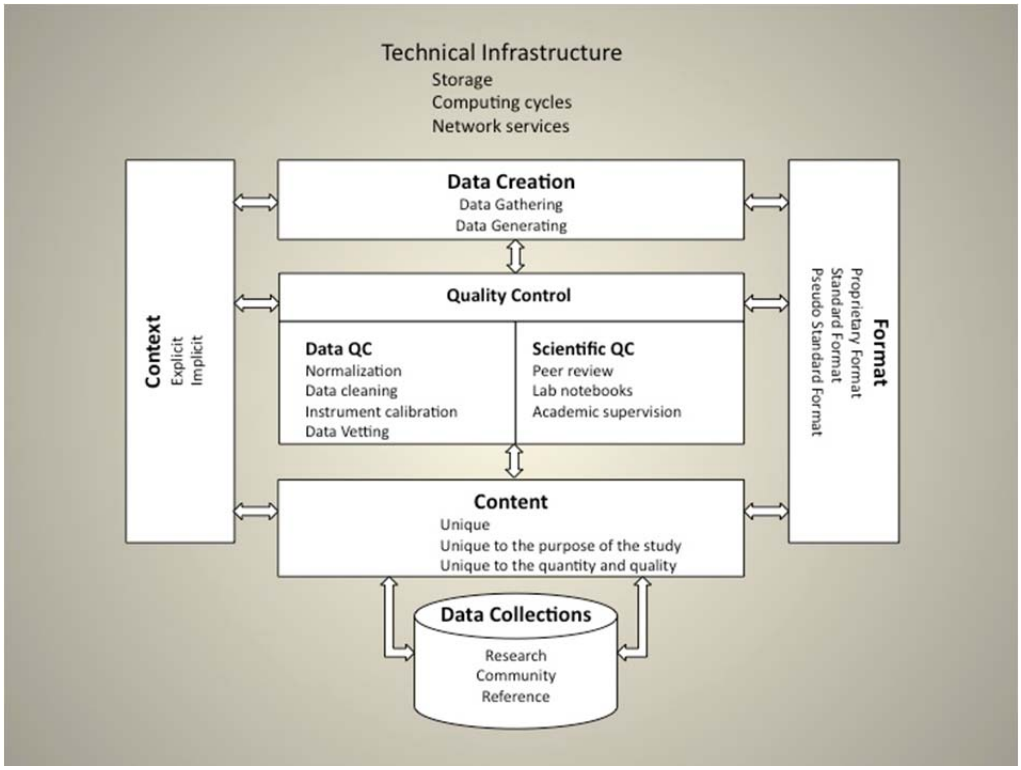


Figure 7. Initial Version of the e-Science Data Environment

The initial model was developed using the grounded theory methodology. This early version of the e-Science Data Environment has 6 components. The four central elements – data creation, quality control, content, data collections – are considered to be a data lifecycle; and context and format are interactive at every step of the lifecycle. Due to the limited number of

researchers interviewed and the nature of the data collected, the interactions were implied and could not be explicitly defined.

The new model of the e-Science Data Environment (see Figure 8) is based on the data from both the preliminary study interviews and the survey responses described in Chapter 5. With this rich set of data, a more detailed model has been constructed.

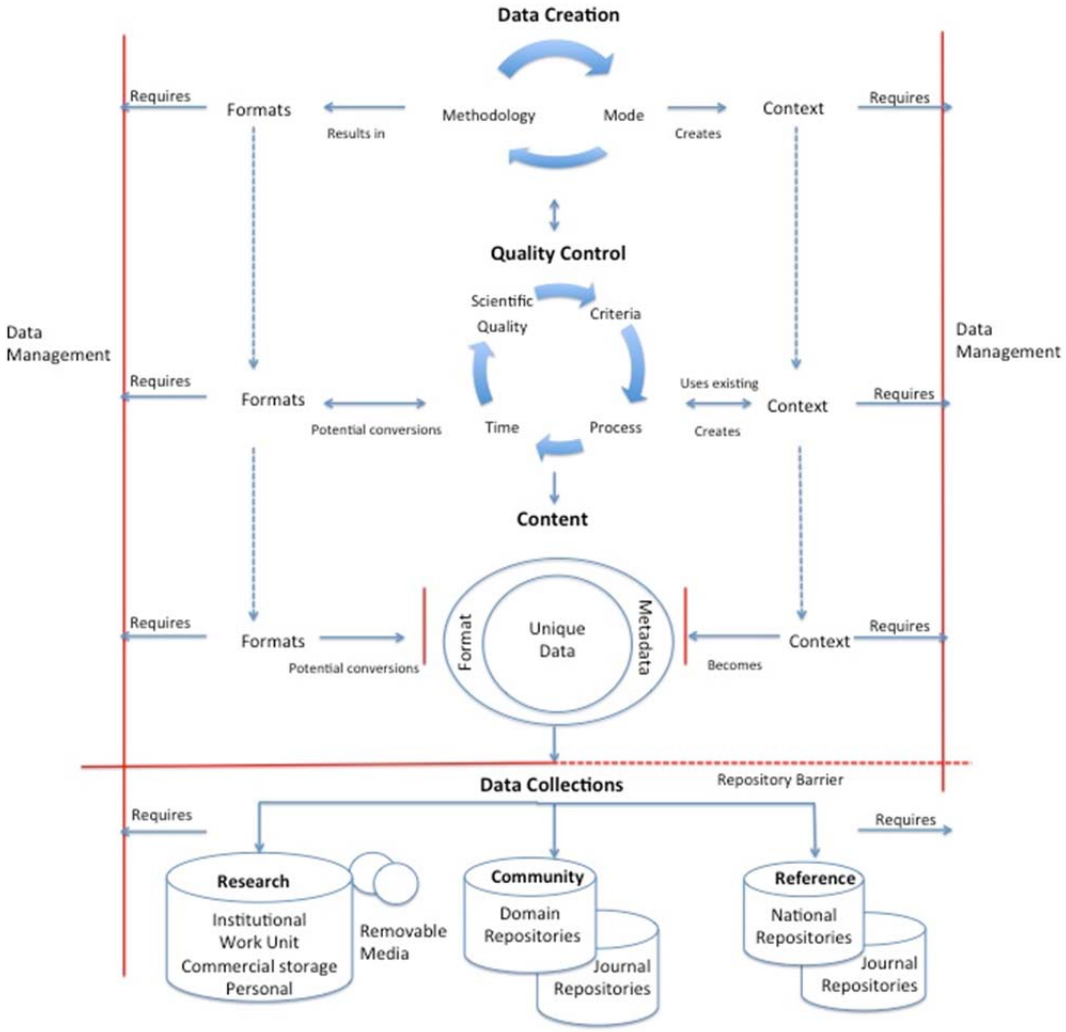


Figure 8. New Model of the e-Science Data Environment

The new model has a more nuanced lifecycle. Data creation and quality control are cycles within the larger lifecycle. Context and data collections have more subcomponents. The interactions with context, format, and data management are described with both function and direction. Barriers to preservation are explicitly addressed (and depicted by red lines).

In the preliminary study, the data about the technical infrastructure was insufficient to build a realistic model. With the data from the current study, the model can be expanded and refined. Rather than belonging to single construct as previously thought, the three components – computing cycles, storage, and data management – are separate. Computing cycles give researchers additional capacity for creating and processing data; however, computation itself has little impact on overall lifecycle or preservation barriers. Storage is an important subcomponent of the e-Science Data Environment and is closely aligned with the final disposition of the data – the data collections construct of the model (see 6.5 for a complete discussion). Data management has become a major component, interacting with all elements of the model; it will be discussed within the context of each component within the lifecycle.

Research is not a linear process. Within a single project, a researcher could have data in any or all of the stages simultaneously. Within each of the stages of the lifecycle, the implications and impacts of the major components change: format in data creation has different implications from format in the quality control stage, which has different implications from format in the final stages. These implications and impacts are explored more fully in each of the following sections.

6.1 Barriers to Preservation

The antecedents to preservations, the actions that data creators need to take to make their data more preservable, are data management, metadata creation, and preservation technologies.

Because these actions are difficult, time consuming, and require expertise or specific resources such as systems and technologies, the antecedents can become barriers to preservation.

6.1.1 Data Management as Barrier

Every step that produces new data or generates context creates a data management event – an opportunity to make decisions on how to name the data file, which format to use to store the data file, where to store the data file, whether to capture content, how to encode the context, whether to create backup copies of this file, and potentially many other decisions. Data management events, these opportunities for decisions that affect preservation, are continually present in the e-Science Data Environment. These decisions are influenced by the institution in which the researcher works, the practices of the particular workgroup or department of the researcher, the domain in which the researcher practices, and the financial resources available to the researcher. Poor decisions, errors in judgment, and lack of knowledge can all lead to data loss. Data that is not available cannot be preserved.

6.1.2 Context as Barrier

Although context and metadata are often considered to be synonymous, for this study, context is considered to be more inclusive than metadata. Context describes the relationships of the data content to its environment (CCSDS, 2002) while metadata is a codified representation of this information, generally using one or more predetermined structured representational formats. Metadata is the method by which data is described so that it can be understood within its context. The process to convert contextual data into metadata remains a barrier to preservation: researchers have more contextual data than can be encoded in their metadata schemas; researchers are not using standardized metadata formats to encode their contextual data, leaving it vulnerable to obsolescence; and researchers are storing their metadata in implicit and

inactionable or in non-standard technologies.

6.1.3 Preservation Technologies as Barriers

Preservation technologies remain a barrier to preservation. As described in sections 2.5.3 and 2.6.3, repository systems are considered to be the primary preservation technologies. In the e-Science Data Environment model, these repository systems are the persistent infrastructure for the final disposition of data described as data collections. It was initially thought that the usability of these systems was the primary barrier to preservation: lack of good tools, significant time investments, and poor interfaces were the primary issues discussed in the literature. This study, however, indicates that the issue is not usability but access. The researchers in this study simply did not use repositories; only a small percentage of researchers indicated that they have deposited data in repository. It is possible that the majority of researchers are unaware of potential repositories for their data. It is possible that researchers are not motivated to contribute data to their domain repositories. It is possible that the researchers do not want to expose their data to community scrutiny. Rather than using repositories to archive their data, the majority of researchers in this study used the least expensive and least preservation-worthy storage technologies to "archive" their data. The technologies used to preserve the research of the great majority of researchers are insufficient for preservation.

6.1.4 Format as Barrier

Format emerged as an additional barrier to preservation. It has been well understood that format is an important factor in preservation; using well known, public, and transparent file formats allows data managers and archivists to process and maintain digital data more easily (Abrams, 2004; Kowalczyk, 2008). In the preliminary study, it was clear that researchers' understanding of standard formats did not conform to the definitions and typologies in the

preservation literature. Researchers include generic syntactic formats such as comma separated values (.csv) and SQL based databases as standard data formats. Standard formats in the preservation literature are defined as both syntactic and semantic formats that are transparent, without copyright or license restrictions, and community-based.

This study shows that the gulf between the two understandings of standard format is greater than previously thought. The researchers in this study described opaque, commercial, proprietary formats from such software packages as Microsoft Office as standard formats. Saving data in these generic or proprietary formats creates a major barrier to preservation, as they require additional context to describe the meaning of the data as well as specialized and often expensive software that needs a specific computing environment to render properly.

6.2 Data Creation

In the e-Science Data Environment, creating data is the first step in the process (see Figure 9). The data creation process has two major components – mode and methodology. The mode is the manner of creation, either generating or gathering. As discussed in section 3.1, data can be *generated* by observations, instruments, or experiments; or data can be *gathered* via databases, vendors, webcrawls, and other processes. Methodology is the set of parameters that defines the processes, practices, and procedures for scientific research. For the e-Science Data Environment, methodology includes such processes as surveys, field studies, case studies, direct observation in experimental situations, analysis of instrument generated data, analysis of existing data sets, modeling and simulation, and text or language analysis.

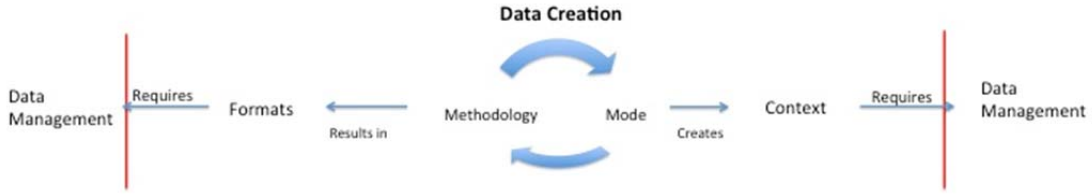


Figure 9. Data Creation Cycle

This research shows that the mode of data creation and the research methodologies interact. The research methodology dictates the mode of data creation; that is, the requirements of the research methodology determine whether the researcher generates new data, uses existing data, or needs a combination of newly generated data along with existing data, as a number of the methodologies use both modes of data creation. Rather than a binary choice of either/or, the mode of data creation is a continuum from exclusively gathering data to exclusively generating data (see Figure 10).

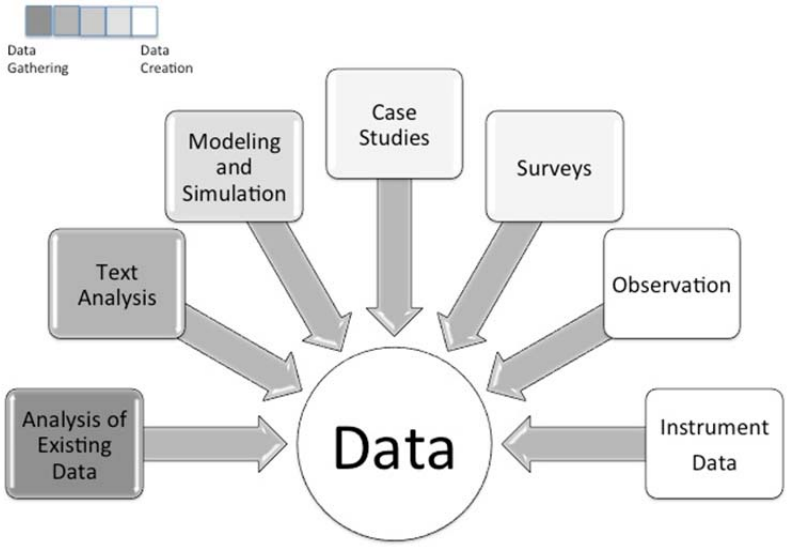


Figure 10. Data Creation Modes and Methodologies

Clearly, analysis of existing data relies on gathering data. This is represented in Figure 10 as dark grey; the color lightens along the continuum toward methodologies that are exclusively data generating. Text analysis almost always involves preexisting textual data in the forms of books, journal articles, newspapers, websites, metadata records, and other content. This textual data must be gathered prior to the analysis. The analysis, then, generates data as the researcher identifies, extracts, indexes, or correlates relevant components. This new data can be incorporated into the text itself using markup languages such as TEI or can be extracted and stored in separate files such as indexes, correlation matrixes, and tag clouds. Modeling and simulation methodologies generate new data but also require existing data. For example, an initial set of existing data is used to seed a model or simulation, which then generates new data, which can then seed further models and simulations. Case studies and surveys can also require both modes of data creation; while primarily data generating methodologies, these methodologies may require some data gathering. In case study methodologies, information about organizations in which subjects participate may be gathered from institutional websites, annual reports, and other external sources. In survey methodologies, data can be generated that needs further explication; for example, in the survey for this dissertation, participants provided their research institution; but the zip codes used to create the participants distribution map in section 5.1.5 were gathered from existing data sources.

Data creation is not a single event in the research process but is an ongoing process throughout the lifecycle. As data is analyzed, additional data is created. This new data can either be ancillary, supporting data or can be the primary research output becoming more important than the original data. For survey data, the analysis data such as the output from statistical programs could be considered ancillary or supporting data as it is used to support

conclusions that are reported in published papers and can be recreated and verified with relative ease. For data that is longitudinal or is merged from many sources, the original data can be less important than the final integrated dataset.

6.2.1 Data Creation and Format

Each methodology has a set of requirements, which often dictates a set of data formats. These formats can be proprietary, syntactic standards, or community-based syntactic and semantic standards. A methodology may require a complex collection of data in multiple file formats of different types. For example, text analysis generally requires one or more input text files in .txt (a syntactic standard) or .xml formats (either syntactic standard or a community standard). The output of the analysis can be additional .txt files, new .xml files, Excel (proprietary formats as either .xls or .xlsx) or .csv files (syntactic standard) for word counts and other simple statistics, or statistical proprietary formats such as .sav for SPSS or .sd*[10] for SAS (both proprietary). Modeling methodologies used in such domains as mesoscale meteorology use sophisticated software with very specific data inputs such as 2- or 3-D terrain data, Doppler wind velocity data, and latitude and longitude data (Pielke, 2002). The format of survey data is dependent upon the software that manages the data. Typical software and formats for surveys are spreadsheets (.csv, .xls or .xlsx), databases (FoxPro, Microsoft Access, and others), and statistical software such as SAS (.sd*) or SPSS (.sav). Observational methodologies produce data in a wide variety of formats such as TIFF and JPEG for microscopy, various MPEG formats for video and audio, and generic types of data formats such as text files and databases. Methodologies that use instruments can generate data in instrument-specific propriety formats as

[10] The * indicating version number of the SAS software: .sd7 for SAS version 7; .sd8 for SAS version 8, etc. (http://support.sas.com/rnd/migration/papers/peaceful.html).

well as community semantic and syntactic standards such as FITS[11] (for astronomical data), NetCDF[12] (for array data), SEG-Y[13] (for seismic data), ROOT[14] (for high energy physics), and FASTA[15] (for protein sequences).

6.2.2 Data Creation and Context

Data sources, instrumentation used, instrumentation settings, experimental conditions, software used, software configurations, and samples used are some examples of contextual information that can result from the process of creating research data. At this stage of the lifecycle, much of this data is implicit, captured in configuration files, lab notebooks, text documents, human subjects testing application forms, and file names. In a small number of domains using specific methodologies, contextual data is captured as data is generated and stored in a standard format.

6.2.3 Data Creation and Data Management

As data is created, data management events involve naming, organizing, saving, and backing up the research data. Determining a standard practice for file naming and file organization is an important task that can help researchers be more efficient and have better control over the research data. Contextual data also needs a set of standard practices for capture and safe storage. The choices made at data creation are among the most important data management decisions in the lifecycle, as they affect the long term preservation of the original research data as well as the contextual data.

---

[11] http://heasarc.nasa.gov/docs/heasarc/fits.html

[12] http://www.unidata.ucar.edu/software/netcdf/

[13] http://pubs.usgs.gov/of/2001/of01-326/HTML/FILEFORM.HTM

[14] http://root.cern.ch/drupal/content/root-files-1

[15] http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml

The potential threats to preservation in the data creation stage are twofold: human error and undetected equipment malfunction. Inadvertent human errors include mislabeling data, misplacing data, and erasing data. Undetected equipment malfunctions can cause data corruption and data loss. Good data management and good quality control practices can mitigate these threats.

6.3 Quality Control

As data is assembled via the multiple modes and methodologies described above, researchers invest significant amounts of time and effort to ensure the quality of their data and of their science. Quality control, the processes by which data is determined to be accurate, complete, and current (Batini & Scannapieco, 2006), is the second step in the e-Science Data Environment model (see Figure 11).

Figure 11. Quality Control Cycle

Scientific quality and quality of data are tightly coupled. As reported above in the preliminary study and confirmed in the current study, researchers overwhelmingly correlated quality control with the quality of their research. Without confidence in the data, there can be no confidence in the results. Determining the criteria by which to judge the quality of the data is a dynamic process that depends on the nature of the project, the nature of the data, and the

specifics of the instruments or the methodologies that were used to create the data. Researchers in domains that use standardized instruments, regularized processes, and quantitative data are often able to develop explicit sets of quality criteria that can be reusable. In domains that are not data intensive or that use qualitative data, researchers are generally not able to or have not perceived the need to develop explicit criteria.

Researchers in "big science" (Weinberg, 1961) – large work groups using large infrastructure, instruments, and equipment, such as biology and geoscience – invest heavily in quality control, spending many hours on data normalization, data cleaning, data integration, instrument calibration, statistical analysis, data modeling, and image processing. Individual researchers (those not a part of "big science" laboratories) were less likely to use these processes. It is possible that this strong divergence in practice between large labs and individuals is due to the scale of the data, the complexity of the data, and/or the nature of group work. A large group of people working with petabyte data from numerous sources requires normalization, integration, and calibration. If the work is distributed among many researchers within the group, the need for data verification process increases. Individual researchers working with data from one source are likely to need none of those processes.

Quality control is a cycle within the larger lifecycle. Scientific quality requirements inform the data quality criteria, which inform the processes that are required to ensure the quality of the data, which takes time and resources, which influences the process of the science. This cycle interacts with the data creation step. As either raw or analyzed data is created via the multitude of methodologies, quality control processes may be performed. These processes are in themselves cyclical and cascading (see Figure 12). Quality control processes can require data format conversions that then generate additional contextual metadata. The number of quality

control processes used increases the number of potential format conversions, the amount of contextual metadata generated, and the amount of time invested.

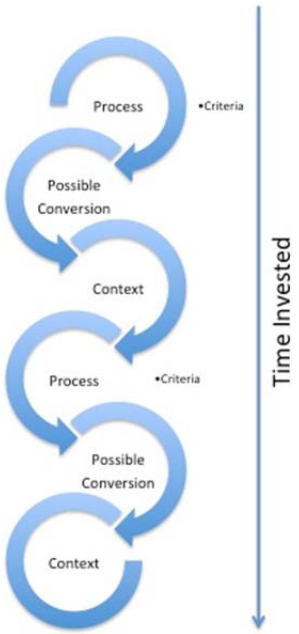


Figure 12. Quality Control Interactions

6.3.1 Quality Control and Format

Quality control processes may require format conversions. The original data files may need to be converted into a new format that the program or process requires, and the output of the program or process may be in yet another format. Examining a simple quality control process can highlight this phenomenon. For example, imagine that a researcher has tabular data stored as a comma separated values (.csv) file and wants to use SPSS to provide descriptive statistics to check for data validity. The researcher would load the .csv file into SPSS, which would automatically convert the data into the SPSS internal database format (.sav). Data can be manipulated, modified, and saved in SPSS, creating a .sav file. As the researcher executes each of the statistical processes, an output file is produced as a .spv file, an opaque, proprietary

format. In SPSS version 19[16], this .spv output file can be exported in any of 8 formats: Excel (.xls), HTML (.htm), Portable Document Format (.pdf), Text – Plain (.txt), Text – UTF8 (.txt), Text – UTF16 (.txt), Word/RTF (.doc), and the original proprietary graphics format (.spv).

Some of the quality control processes are ends to themselves; that is, the assessment of quality is captured in the output from these processes and is not used in further processes. That is likely to be the case in the simple example above. Other quality control processes are part of a series of steps that cleans, massages, augments, filters, and collates data. These processes in series can create a number of files in different formats, each of which can require a conversion to another format for further processing or into the final format

6.3.2 Quality Control and Context

Contextual data is generated throughout the quality control cycle. This contextual data includes quality control criteria, domain and workgroup norms, processing algorithms, determinations of statistical outliers, and a wide variety of other decisions, both explicit and implicit. The amount of contextual data generated is directly related to the quality control processes. As the number of quality control processes increases, the amount of contextual data increases. This is only logical. Each process has a motivation (remove outliers, reconcile scale), a specific set of rules (remove those data points that are greater than a specific standard deviation or convert from zip code level data to state level), and an instantiated implementation (a set of parameters for statistical program such as SAS or SPSS, a program developed for this specific purpose, a Schematron[17] plugin to an XML editor), each of which produces and/or contains contextual data.

---

[16] http://www-01.ibm.com/software/analytics/spss/ (copyright 2010)

[17] Schematron is an XML language used to validate encoding and to find patters for automatic markup. See http://www.schematron.com/overview.html

Much of this data is implicit, stored in software configuration files, software source code, directory structures and file names, lab notebooks, documentation, and other text documents. Some of the data is explicit, stored in databases or spreadsheets. For a small number of researchers in specific domains, data is stored in a community syntactic and semantic standard format.

6.3.3 Quality Control and Data Management

Quality control processes can create multiple new files in a variety of formats, all of which require data management. Determining which of these files to keep is a crucial data management decision. As was noted in the preliminary study and confirmed by comments in the survey, researchers lack confidence that they know which files will be important over time.

Maintaining control over the multiple versions of files produced in quality control processes is a function both of data management and context management; it involves understanding and documenting the relationships between revised, derivative, and/or intermediate datasets. As a function of data management, version control requires that the datasets have clear organization and file naming as well as the obligatory safe storage and adequate backup. As a function of context management, version control requires that the transformations from an original dataset to a derivative file to an intermediate processed file are documented and that this documentation is available to the data manager. As the data is managed primarily by researchers and/or graduate students, data management at this level can be a significant barrier to preservation. The complexity, the amount of time required, the lack of tools to automate these data management tasks, and lack of standard practice increase the probabilities of errors.

At this stage in the lifecycle, threats to preservation involve deliberate but mistaken data deletion, inadvertent human errors, mislabeling data, misplacing data, software obsolescence, undetected data corruption introduced from quality control processes, and equipment malfunctions.

6.4 Content

Through the research process, data becomes content: it has value; it has form; it has meaning. That is, for data to become content, it must have value; it must be unique in some way that extends human knowledge, that adds value to the scientific record (see Figure 13).

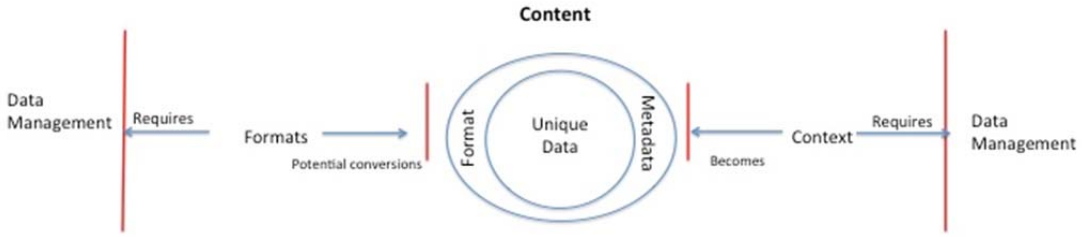


Figure 13. Data to Content Transformation

For data to become content, it must have a form that is readable, renderable, and usable. For data to become content, it must be understood within its context; it must have metadata that describes its meaning. As with the other stages in the e-Science Data Environment, this is not a single event step. The transformation from data to content can be complete for some data while other data is still in the data creation state and yet other data is in the quality control stage.

Within the e-Science Data Environment, uniqueness is the measure of the value of data. Uniqueness has been the primary assessment criterion for data preservation (see section 4.1.3). The overwhelming majority (85%) of researchers in the study indicated that they consider their

data to be unique. This uniqueness can be due to the nature of the data or to the value added to data through the research process. Data can be unique because of the nature of the data; that is, the data has been created by a unique process either through observation or experiment that has never occurred previously. This is the traditional view of data uniqueness that has been the criterion for preservation assessment: data that cannot be computationally recreated.

Researchers, both in the preliminary study and in the current study, indicate that uniqueness, value, can be attributed to the contributions of the research process; that is, researchers can add value to existing data that makes the data unique. Research that creates longitudinal data, research that collects and collates data from different sources to create a new and integrated dataset, research that integrates analysis into the data (such text encoding and geocoding), and research that adds context to existing data all add value to data; these processes create data that is considered by the researchers to be unique. This data cannot be easily recreated; the processes that create this data cannot be easily rerun.

6.4.1 Content and Format

All throughout the e-Science Data Environment, the construct of format, the representational expression of data, is cumulative; that is, the format decisions made at the beginning of the research cycle have implications at the later stages. As the data becomes content, many of the decisions made previously have implications for the longevity and the re-usability of the content.

Researchers in this study overwhelmingly use undocumented or proprietary standards. That is, the most frequently used formats are either generic syntactic formats with no internal semantics to describe the data such as comma separated values format (.csv) or are formats that are opaque, are commercially owned, and have strong commercial software dependencies (such

as Microsoft Office formats for Word or Excel). Both of these types of formats present significant barriers to preservation. Generic syntactic formats have obvious advantages during the research process, as they are flexible, allowing virtually unlimited numbers and types of fields and do not require specific content typing such as date formats, controlled vocabularies, and decimal precision. These flexible features that are so useful in research mean that the context for the data content is external to the data; the meaning of individual fields, the rules by which these data elements were created, are not captured semantically within the file. These generic syntactic formats create barriers to preserving the meaning of the data.

The commercially owned, opaque formats can also provide obvious benefits during the research process. The commercial software packages used to create data in these formats have powerful features such as automatic replication of data, strong automatic data typing, highly useful built-in functions such as statistical formulae, data visualization, spell checking, tables, embedded images, and many others. However, many of these features produce data in proprietary internal structures. These opaque, proprietary formats create barriers to preservation as the data is locked in undocumented formats that require specialized, commercial software to render and process. The software may be part of ubiquitous computing platforms that are available on every computer used by the researchers; thus the researchers assume these proprietary formats will always exist and be available.

6.4.2 Content and Context

In order for data to be transformed into content, its meaning, its relationship to its environment, must be codified; that is, the contextual data that was generated throughout the process must be processed into metadata. This process of generating metadata from context is a transformation in itself, creating a structured representational format from scattered bits of

information. This transformation is an imperfect process. This research shows that researchers face major obstacles to creating preservation quality metadata: finding appropriate standard metadata formats, mapping the contextual data, and allocating the resources required to create the metadata.

In general, researchers do not use domain- or community-developed semantic and syntactic metadata standards; less than 5% of researchers in this study report using such standards. The possible reasons for this low adoption rate of standards are many: it is likely that many domains do not have standard metadata formats; it is possible that researchers are unaware of existing metadata standards; it possible that the specific research of participants in this study do not fit the existing standards; it is possible that the lack of metadata tools prohibit adoption; or the effort to use an existing standard exceeds perceived benefits. The lack of standard metadata format use is a significant barrier to preservation.

For a small number of researchers in domains that use specific instrumentation and standards, contextual data is captured as data is generated and stored in the community standard format. For these researchers, creating metadata is not the barrier to preservation that others face. However, if these researchers use multiple methodologies or instruments that do not produce standard output, they will have the same issues as other researchers: identifying, locating, and deciphering their contextual data to create metadata.

Although researchers are not using community-based syntactic and semantic metadata standards, they are creating metadata. Because a majority of researchers in this study use explicit and actionable metadata that is stored in databases and spreadsheets, it must be concluded that they have some type of metadata scheme. This scheme could be idiosyncratic,

for the purpose of a specific project or type of research; it could be a lab-based specification; or it could be a subset or a superset of a community standard.

Although the researchers seem to be developing their own metadata schemes, they report having more contextual data than they can encode. Thus, transforming context to metadata remains a barrier even when researchers use their own metadata schemes. For many researchers, there is no clear set of steps or process from context to metadata; creating metadata is a manual process of gathering the contextual data from the variety of locations and mapping the context to the metadata scheme. The contextual data is analyzed to determine where in the metadata scheme this data should be stored. Mapping the contextual data into metadata formats requires a full understanding of the data creation and the quality control processes as well as the location, the type, and meaning of the contextual data. As more methodologies and quality control processes are used, the amount and the complexity of the contextual data increases as do the difficulties of mapping that contextual data to metadata schemes.

The resources required to create the metadata are substantial. The lack of such resources is a significant barrier to preservation. Quantifying the amount of time required to create adequate metadata is difficult; the amount of time and resources required to create metadata is dependent on the type of project, the amount of data, the types of data, and the number of different types of data. As the complexity of the context increases, the amount of time required to create the metadata increases. For some research projects, the amount of effort is measured in multiple months of effort or in numbers of full time staff. Funding for additional resources for metadata creation is sparse. In general, researchers are reluctant to spend their research funds on metadata specialists as they have concerns about their ability to use the specialists effectively,

having sufficient work to keep a full time person occupied, and the cost of transferring complex, domain specific knowledge.

6.4.3 Content and Data Management

Content is the result of good data management throughout the lifecycle, when the data and the context are available, are knowable, and are viable. The decisions made during the previous steps are now visible. Are all of the necessary data files available? Are they intact without errors? Are the contextual data files (both digital and analog) sufficient, available and viable? Are the appropriate intermediary files available? Is the researcher assured that these files are the "right" ones, the most current ones, the most accurate ones? If the researcher can answer these questions affirmatively with confidence, the data management has not been a barrier to this point. However, the data management barrier has not been eliminated. Data management will be an ongoing process over time even as the data becomes part of a data collection.

6.5 Data Collections

Making arrangements for the final disposition of data is the last step of the e-Science Data Environment model; while this may invoke a funereal vision of buried data, the process is focused on providing ongoing access to data. As defined by the Merriam-Webster dictionary (2011), disposition means an orderly arrangement or the transfer of administrative control to another. These are the primary options available to researchers – to maintain their own data, hopefully in an orderly manner, or to transfer responsibly to another entity. Within the e-Science Data Environment, the construct of data collections describes a taxonomy of the final disposition of research data: research collections, community collections, and reference collections (NSB, 2005) (see Figure 14). Research data collections refer to the output of a single researcher or lab

during the course of a specific research project. Community data collections generally serve domain or other well defined area of research. At the highest level, reference data collections are broadly scoped, widely disseminated, well funded collections that support the research needs of many communities (NSB, 2005).

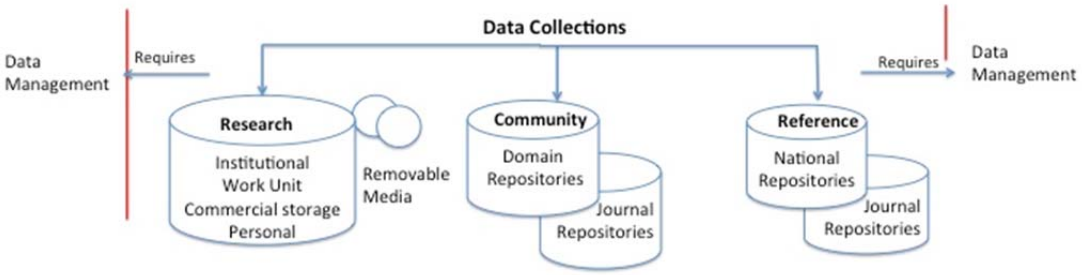


Figure 14. Data Collections

While some researchers do nothing to their data once a project is complete and the data remains as it was, most researchers must take an action to free workspace within their computing and storage infrastructure, such as copying files. Very few researchers decide to delete data; those who do are generally pruning the data, removing intermediate results files that they considered to be unimportant.

Research collections are by far the most common disposition of data. Research collections are primarily supported by the individual researcher on removable media such as media as CDs, DVDs, or hard drives. They maintain the responsibility for their data, not by choice but by necessity; they have no better options. As well, research collections can be supported in a lab or work group environment. Individual researchers are able to transfer the control of their data to a lab supported data archive. The research lab takes responsibility for managing the data storage environment but generally does not commit to long term preservation. Institutional data archives can also support research collection. These archives take responsibility for data management,

long term preservation, and ongoing access. A small number of researchers have begun to use commercially available storage services such as Amazon and Google for their research collections. These services provide a stable technology base that provides online access for a very low cost.

Community or reference data collections serve as the final data disposition for some researchers. Researchers have a variety of motivations for contributing their data to these collections. According to the results of this survey, funding agency mandate is the primary motivation for researchers to contribute to both community and reference collections. For researchers contributing to community collections, personal initiative is the second most important motivation followed by standard practice in research work group. For research collections, journal mandates are the second most important motivation followed by standard practice in their research group.

6.5.1 Data Collections and Format

In the e-Science Data Environment, format conversions occur throughout the research process. Scenarios include conversions from a single source to a single standard format, conversions from a single source into multiple standard formats, conversions from multiple sources into a single standard format, and conversions between multiple intermediate formats into a final standard format. The type of data collection determines the future format conversion scenarios. For research collections, conversions to a new format or upgrades to a newer version of a format will occur when the data is being accessed for reuse. That is, data in research collections will remain in the existing formats until a new use for this data forces conversion. For community and reference collections, future format conversions will be the responsibility of

that data collection. The repository managers and the community will determine the necessity of format changes, upgrades, and conversions.

6.5.2 Data Collections and Context

Creating metadata and depositing the data in repositories have been seen as barriers to preservation. There have been concerns with the usability of the repository systems, the lack of tools, and the amount of effort required to create metadata and to deposit data. Repositories for reference and community data collection have different metadata requirements from institutional, journal, and commercial repositories. Reference and community data collections generally have a higher requirement for metadata, specifying specific contextual information in specific formats. Institutional repositories generally require minimal bibliographic information (such as creator, title, and abstract). Journal repositories generally assume the published paper to contain sufficient contextual information. Commercial repositories require only a file name and administrative information such as the researcher's name, email, and billing information.

For all types of repositories, researchers in this study found deposit processes to be easy to use. This implies that the amount of time and effort was reasonable for these researchers. However, the reference and community data collection repositories used by participants in this study are heavily concentrated in the biological and geosciences, many of which use data formats with integrated metadata; that is, the contextual data is embedded in the data format when the data is created. It is possible that the nature of the data that these repositories store makes the deposit process less burdensome. It is also possible that motivations to deposit influence perceptions of ease of use. Researchers who have a strong research culture of repository use or those who have strong personal motivations to use data repositories may have a higher threshold

of patience, thus rating deposit processes as easier than those who are externally motivated to use repositories.

6.5.3 Data Collections and Data Management

As with the other components of the data collection construct, data management is bifurcated: issues with research collections are different from the issues with community and reference collections. The repository services of community and reference data collections assume the responsibility for data management upon data ingest. The technology and data management practices of these repository services are generally opaque, and thus claims of preservation services are difficult to verify (Kowalczyk & Shankar, 2011). Despite the opacity of the services, researchers are using these services with the expectation of persistent archiving and preservation. The threats to these collections include loss of funding, equipment malfunction, equipment obsolescence, undetected data corruption, and human error.

Research data collections, the output of a single researcher or lab, have ongoing data management requirements such as regular backups, offsite backup storage, and ongoing documentation. The probability of long term data management for research collections is low when the ongoing responsibility lies with an individual researcher or graduate student. Individual researchers are focused on their current research, and graduate students are short term resources with little motivation or few opportunities for knowledge transfer of data management requirements to the next cohort. When an individual researcher has department-level support for data management, the probability for ongoing archiving of the research data collection increases but only for the duration of the researcher's tenure. If or when a researcher leaves the department, the data management tasks once again become the responsibility of the researcher, assuming that the researcher has retrieved and taken possession of the data. Individual

researchers may have an option to deposit their data into an institutional repository. An institutional repository often has some characteristics of a community collection, such as ongoing data management; however, the researcher cannot presume that data format migration will be part of an institutional repository's services. That responsibility may still remain with the researcher. Research data collections of large labs may have additional data management resources available: more funding for staff, additional graduate students, or more funding for storage. In large labs, data management can be the responsibility of the group rather than the individual researcher. Thus data maintained by large labs may have a higher probability of persistence over time.

In addition to the threats of funding, equipment malfunction, equipment obsolescence, undetected data corruption, and human error as noted for community and reference data collections, research data collections are threatened by software obsolescence, lost media, and deliberate but mistaken data deletion.

6.6 Limitations of this Study

This study is intended to be a broad-based survey of research data practices. The sample frame was based on recent National Science Foundation grant awardees. Although there was broad-based participation across geographic area, institutions, and domains, the sample was skewed toward high level, established researchers in the U.S. The primary attempt to broaden the participation, a request to pass the survey on to others, was less than successful; less than 6% of respondents were outside the original sample list. Having a sample with researchers with different roles, such as graduate student or post-doc, would provide a more balanced view and perhaps more generalizable results.

6.7 Future Work

The e-Science Data Environment is an emerging model that can help researchers, librarians, archivists, and data managers understand both the antecedents of and the barriers to preserving scientific research data. The model was developed with data from researchers primarily from the sciences. It is likely that the model could be generalized to include all types of e-research; surveying researchers in the humanities and social sciences will be the next step.

There remain gaps in the e-Science Data Environment, particularly in understanding the process of creating metadata from contextual data and the requirements and processes of format conversion. This survey confirms the concept of context as a superset of metadata; that is, metadata does not capture all of the information about research methodologies, quality control processes, and scientific goals that is created. However, the process of creating metadata, encoding the context into a formatted digital representation, was not fully explored in this study. It is not clear how researchers decide on a metadata format. Nor is it clear how researchers map their contextual data into their metadata formats. As metadata creation remains a barrier to preservation, understanding more about the metadata creation process is important.

This study revealed that researchers frequently convert data between different formats during the research process. However, the specific criteria and requirements for these conversions were not explored. Understanding the motivations for conversions may allow developers of research tools to make these transformations easier and, perhaps, less frequent. Future work includes developing a study to extend the e-Science Data Environment to describe more specifically when formats conversions occur during the research cycle: understanding the frequency and nature of format conversions required during the data creation process, during the quality control process, during the transition from data to content to final disposition.

The data collected for this dissertation is rich and has many stories to tell. Additional analysis of this data could produce new insights and develop new models. It would be possible to analyze this data to develop models of practice by domain, which could help community data collection repository developers and managers refine their functionality and find new opportunities to participate in the research process. Another analysis of the data could provide a view of institutional-level research environments; looking closely at the data by institution and triangulating with external sources on policies and incentives, a model of research support could emerge. With additional data via interviews, a more robust understanding of the end-of-life and final disposition of data could be developed. One aspect of final disposition of data is the emergence of journals as data repositories. There is little understanding about the relationships between community data collections and journal data repositories, and questions abound: are they complementary or competitors; what is the mission of the journal data repositories; what are the economics of journal data repositories? Using case study methodologies to gather additional data, an examination of data management practice in labs and data management events could produce a stronger model of the research environment.

7 Conclusions

In Chapter 2, four research questions were posed concerning the research practices of scientists, the lifecycle of research data, and the antecedents, barriers, and threats to preservation. This research used a mixed methodology approach to answer the questions: a preliminary study to develop a data lifecycle model of research data using grounded theory with a theoretical sample of polar opposite case studies and a broad-based survey that produced both qualitative and quantitative data which was used to expand and generalize the lifecycle model. The results of this research have provided a number of insights that can significantly enhance the understanding of preserving research data.

7.1 Modeling the Research Data Lifecycle

This work integrates prior research in digital preservation, computer science, information science, and domain sciences. The major contribution of this research is the development of the e-Science Data Environment, a data lifecycle that provides a generalized model of the research process in science. Lifecycles provide an important and useful framework for understanding data preservation (Beagrie, 2006; Rice, 2007; Rumsey, 2010). As noted in section 2.7, many of the extant lifecycles are either completely generic or based on a very narrow domain. Having a broad-based, generalized model of the scientific data lifecycle based on a large survey of researchers from multiple domains provides a theoretical basis to explain and predict both the antecedents and barriers to preservation.

This research has identified a set of antecedents to preservation. These antecedents – data management, contextual metadata, and preservation technologies – can become barriers to

preservation when researchers do not have access to appropriate resources. Data management, a set of skills and technologies required to ensure the safe keeping of data, is primarily the responsibility of the individual researcher, as institutions do not generally provide data management support to researchers. Creating metadata from the contextual data is a time-consuming task that is inadequately staffed and funded; however, researchers are unconvinced that external resources such as data librarians would help. Researchers are concerned about both the expense of domain knowledge transfer and the effort to manage the workflow. Preservation technologies are not used frequently by the researchers in this study. The cause of this low use was not explored in this study but warrants further investigation. Although the responsibilities for the antecedents to preservation rest primarily with the researchers, institutions and funding agencies can develop policies, services, training, and systems to encourage preservation as data is created.

In addition to the three antecedents and barriers of data management, contextual metadata, and preservation technologies, a new barrier to preservation has been exposed via this research; the use of non-standard file formats. As discussed in detail below in section 7.2.3, researchers in this study do not use syntactic and semantic community or domain data standard formats, making their data more difficult to use and preserve over time.

7.2 Research Practices

By creating quantitative measures, this study provides specific, numeric descriptions of current research practices including data quality control, categories of the uniqueness of data, file formats, the final disposition of data. Many of the results of this study expand the current understanding of research practices.

7.2.1 Data Quality Control and Scientific Quality

This study showed that there are two very distinct understandings of quality: the quality of the science and the quality of the data. The scientific process has a well established quality control mechanism in peer review. Data quality has no such established, predictable, and vetted process. Data quality control is often an *ad hoc* set of processes designed to ensure that the original data is correct as well as normalizing the data to allow accurate merges from disparate sources and to reconcile different scales. There is growing concern that data quality is not as fully transparent or integrated into the peer review process as it should be to validate the quality of the science.

7.2.2 Uniqueness

An important contribution of this research is the reevaluation of the preservation assessment paradigm of uniqueness. The literature indicates a binary judgment of uniqueness: data is unique and should be preserved; or data is derived, can be recreated, and need not be preserved (Gray et al., 2002; Henty et al., 2008; Key Perspectives, 2010; Lord & McDonald, 2003; Lyon, 2007). The results of this research study show that uniqueness is more complicated than previously thought. Scientists in the study described multiple ways in which data could be considered unique. The first is that the nature of the data is unique: the data is of an observation of a singular nature or the data is from an experiment with an exact preparation, processing, and scientific goal. Data can also be unique because additional value was contributed by the researcher through the scientific process; that is, the data is unique because of the quantity and quality of the data, the level of uniformity and integration of the data, the breadth of data, longitudinal nature of the data, the integration of analysis within the data, and the added value of metadata. The researchers disagree with the proposition that the processing to create uniformity

or integration is simple computation. They perceive that their data is unique *because* of the processing, normalizing, merging, and cleaning. Much of this work is done by hand, requires intellectual input, and becomes irreplaceable.

7.2.3 File Formats and Standards

File formats for both research data and contextual metadata are a major component of the e-Science Data Environment. Format is the structured representation of the data and is used by programs to ready, process, and render the data. Without knowledge of the file format, the data is unusable. Thus, format is a significant predictor of the potential for preservation.

Only a small percentage of researchers use syntactic and semantic domain- or community-based standard file formats for their research data and/or contextual metadata. Many researchers view widely available, commercial, opaque, and proprietary file formats as standard. The ubiquity of these formats misleads researchers into assuming their long term viability and availability. Other researchers confuse generic computing syntactic standards such as comma separated values (.csv), SQL based databases, and text encoding standards (ASCII) as standard formats. These formats do not have embedded data elements descriptors, the lack of which results in the inability to understand the meaning of the data and to process the data; without accompanying documentation about the internal structure and meaning of the elements, the data is useless. This research shows that lack of syntactic and semantic domain or community based standard file formats is a significant barrier to preservation.

7.2.4 Data Collections and the Final Disposition of Data

This research has provided new insight into the final disposition of research data. Most research data is not maintained in preservation quality technologies. Rather, research data is stored primarily on removable media, such as CDs, DVDs, and external hard drives, or on

departmental or lab servers. A small percentage of researchers contribute their data to community or reference collections that include commercial journal data repositories, open source journal data repositories as well as the more traditional data collections.

7.3 Conclusion

Curating, preserving, and providing access to scientific data is vital to the health of the scientific enterprise (Iwata, 2008; Schofield, Bubela, Weaver, Portilla, Brown, Hancock, et al., 2009; Rusbridge, 2007). This dissertation develops a new theoretical model for describing the lifecycle of research data that accounts for both the antecedents and barriers to preservation. This research provides new insights into the workflow of digital science process by quantifying the current state of digital science data management and describing the environments in which data is created, used, saved, and preserved.

8 References

Abrams, S. A. (2004). The role of format in digital preservation. *Vine*, *34*(2), 49–55.

Aktaş, M. S., Fox, G., & Pierce, M. (2005). Managing dynamic metadata as context. *Proceedings of the 2005 International Computational Science and Engineering Conference (ICCSE)*. Istanbul, Turkey: IEEE. Retrieved from http://www.opengrids.org/hybrid/publications.html

Altman, M., Adams, M. O., Crabtree, J., Donakowski, D., Maynard, M. M., Pienta, A., & Young, C. H. (2009). Digital preservation through archival collaboration: The data preservation alliance for the social sciences. *American Archivist*, *72*(1), 170-184. Retrieved from http://archivists.metapress.com/content/EU7252LHNRP7H188

Anciaux, N., Van Heerde, H., Feng, L., & Apers, P. (2006). Implanting life-cycle privacy policies in a context database. *Centre for Telematics and Information Technology.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.7294&rep=rep1&type=pdf.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *16*(7). Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Anderson, W. L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal*, *3*(30 December 2004), 191-202. Retrieved from http://www.jstage.jst.go.jp/article/dsj/3/0/191/_pdf

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

Atkins, D. (Ed.) (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report on the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure.* National Science Foundation: Arlington, VA. Retrieved from http://www.nsf.gov/od/oci/reports/toc.jsp.

Baker, M., Keeton, K., & Martin, S. (2005). *Why traditional storage systems don't help us save stuff forever.* Technical Report 2005-120. Palo Alto, CA. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.2375&rep=rep1&type=pdf.

Barateiro, J., Antunes, G., Cabral, M., Borbinha, J., & Rodrigues, R. (2008). Using a Grid for digital preservation. In G. Buchanan, M. Masoodian, & S. J. Cunningham (Eds.), *Digital Libraries: Universal and Ubiquitous Access to Information* (5362), 225-235. Berlin: Springer. Retrieved from http://www.springerlink.com/content/k71v8x6081738x18.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., et al. (2009). NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Research*, *37*(Supplement 1 Database), D885-D890. Retrieved from http://nar.oxfordjournals.org/cgi/content/full/37/suppl_1/D885

Batini, C., & Scannapieco, M. (2006). *Data quality: Concepts, methodologies and techniques*. New York, NY: Springer.

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

Beagrie, N., Beagrie, R., & Rowlands, I. (2009). Research data preservation and access: The views of researchers. *Ariadne, 60*. Retrieved from http://www.ariadne.ac.uk/issue60/beagrie-et-al/

Bell, G., Hey, T., & Szalay, Alex. (2009). Beyond the data deluge. *Science*, *323*(5919 6 March 2009), 1297-1298. Retrieved from http://www.cloudinnovation.com.au/Bell_Hey _Szalay_Science_March_2009.pdf

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* (pp. 269 - 277). Vancouver, BC, Canada: ACM. Retrieved from http://portal.acm.org/citation.cfm?id=1255175.1255228

Borgman, C., Wallis, J., & Enyedy, N. (2007). Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, *7*(1-2), 17-30. doi:10.1007/s00799-007-0022-9

Borgman, C.L., Bowker, G.C., Finholt, T.A., Wallis, J. C. (2009). Towards a virtual organization for data cyberinfrastructure. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (pp. 353-356). Austin, Texas:ACM & EEE. Retrieved from http://portal.acm.org/citation.cfm?id=1555400.1555459

Borgman, Christine L. (2007). *Scholarship in the Digital Age: Information, infrastructure, and the Internet*. Cambridge, MA: The MIT Press. Retrieved from citeulike-article-id:6338113

Bose, R., & Frew, J. (2005). Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys (CSUR)*, *37*(1), 1-28. Retrieved from http://portal.acm.org/citation.cfm?id=1057977.1057978

Brown, A. (2003). Selecting storage media for long-term preservation. *Digital Preservation Guidance Note 2, Issue 1* (Vol. 2). London. Retrieved from http://www.nationalarchives.gov.uk/documents/selecting_storage_media.pdf

Brown, A. (2006). Automatic format identification using PRONOM and DROID. *Digital Preservation Technical Paper 1, Issue 2*. London: National Archives of the United Kingdom. Retrieved from http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf

Buneman, P., Abiteboul, S., Szalay, A., & Hagehülsmann, A. (2006). Laying the ground: Semantics of data. *Towards 2020 Science* (p. 15). Cambridge, England: Microsoft Corporation. Retrieved from http://research.microsoft.com/towards2020science/downloads.htm

Cannataro, M., Conguista, A., Pugliese, A., Talia, D., Trunfio, P., & Congiusta, A. (2004). Distributed data mining on grids: Services, tools, and applications. *IEEE Transactions on Systems, Man, and Cybernetics*, *34*(6), 2451 - 2465. doi:10.1109/TSMCB.2004.836890

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, CA,: Sage Publications.

Cheung, K., K., Hunter, J., Lashtabeg, A., & Drennan, J. (2008). SCOPE: A scientific compound object publishing and editing system. *The International Journal of Digital Curation*, *3*(2), 4 - 18. Retrieved from http://ijdc.net/index.php/ijdc/article/viewFile/84/55

Chin, G., Jr. & Lansing, C. S. (2004). Capturing and supporting context for scientific data sharing via the biological sciences collaboratory. *Proceedings of the 2004 ACM conference on Computer Supported Cooperative work* (pp. 409-418). Chicago, Illinois, USA: ACM. doi:10.1145/1031607.1031677

Choudhary, A., Kandemir, M., No, J., Memik, G., Shen, X., Liao, W., Nagesh, H., et al. (2000). Data management for large-scale scientific computations in high performance distributed systems. *Cluster Computing*, *3*(1), 45 - 60. Retrieved from http://dx.doi.org/10.1023/A:1019063700437

Churchill, G. A., & Churchill Jr., G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*(1), 64-73. Retrieved from http://www.jstor.org/stable/3150876

Coles, S., Carr, L., & Frey, J. (2007). *Final Report: The Repository for the Laboratory*. Southampton, England: Joint Information Systems Committee (JISC) Digital Repositories Programme and the University of Southampton,. Retrieved from http://ie-repository.jisc.ac.uk/166/1/9R4Lfinalreport.pdf

Committee on Data for Science and Technology. (2002). *CODATA Workshop on Archiving Scientific & Technical (S&T) DATA Report* (p. Section 3.2.1). Pretoria, South Africa, May 20-21: South African National Committee for CODATA, CODATA Working Group on Data Archiving and the National Research Foundation of South Africa. Retrieved from http://stardata.nrf.ac.za/Codata/CodataReport_2002.pdf

Consultative Committee For Space Data Systems. (2002). *Reference model for an open archival information system (OAIS), recommendation for space data system standards. CCSDS 650.0-B-1. Blue Book.* Washington, DC: National Aeronautics and Space Administration. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf.

Crow, R. (2002). *The case for institutional repositories: A SPARC position paper.* The Hague, Netherlands: The Scholarly Publishing & Academic Resources Coalition. Retrieved from http://ignucius.bd.ub.es:8180/dspace/bitstream/123456789/315/1/Crow_02.pdf

Dasu, T., Vesonder, G. T., & Wright, J. R. (2003). Data quality through knowledge engineering. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)* (pp. 705-710). Washington, DC, August 24-27: ACM. doi:10.1145/956750.956844

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

Dey, I. (2007). Grounding categories. In A. Bryant & K. Charmaz (Eds.), *The Sage handbook of grounded theory* (pp. 167 - 190). Thousand Oaks, CA: Sage Publications.

Disposition. (2011). In Merriam-Webster's collegiate dictionary. Retrieved from http://www.britannica.com.ezproxy.lib.indiana.edu/

Ecological Society of America. (2011). ESA Data Registry. Retrieved 2011, from http://www.esapubs.org/archive/archive_D.htm.

Eisenhardt, K M. (1989). Building theories from case Study Research. *The Academy of Management Review*, *14*(4), 532-550. Retrieved from http://www.jstor.org/pss/258557

Fendt, K. H. (2004). The case for clinical data quality. *Data Basics, the publication of the Society for Clinical Data Management*. Retrieved from http://www.dqri.org/papers/download/case_data_quality.pdf

Frew, J., & Bose, R. (2001). Earth system science workbench: A data management infrastructure for earth science products. *Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001*. Fairfax, VA, July 18-20: IEEE Computer Society. doi:10.1109/SSDM.2001.938550

Friedman, K. (2003). Theory construction in design research: criteria: approaches, and methods. *Design Studies*, *24*(6), 507 - 522. Retrieved from http://www.sciencedirect.com/science/article/B6V2K-49JHJTM-1/2/1388febf702daeb85c00d0ef51080e4f

Galloway, P. (2004). Preservation of digital objects. In B. Cronin (Ed.), *Annual Review of Information Science and Technology, Volume 38* (pp. 549–590).

Gentil-Beccot, A., Mele, S., Holtkamp, A., O'Connell, H. B., & Brooks, T. C. (2009). Information resources in high-energy physics: Surveying the present landscape and charting the future course. *Journal of the American Society for Information Science and Technology*, *60*(1), 150 - 160. Retrieved from http://dx.doi.org/10.1002/asi.20944

Gershon, D. (2002). Dealing with the data deluge. *Nature*, *416*(6883), 889-891. Retrieved from http://dx.doi.org/10.1038/416889a

Gladney, H. M. (2004). Trustworthy 100-year digital objects: Evidence after every witness is dead. *ACM Transactions on Information Systems*, *22*(3), 406-436. Retrieved from http://doi.acm.org/10.1145/1010614.1010617

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research* (p. 271). Chicago: Aldine Publishing Company.

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. Retrieved from http://doi.acm.org/10.1145/1107499.1107503

Gray, J., Szalay, A. S., Thakar, A. R., Stoughton, C., & vandenBerg, J. (2002). *Online Scientific Data Curation, Publication, and Archiving* (Vol. 200). Redmond, WA: Microsoft Research. Retrieved from http://arxiv.org/abs/cs.DL/0208012

Green, A. (2008). *Data Documentation Initiative DDI*. *DataShare*. Edinburgh, Scotland. Retrieved from www.disc-uk.org/docs/DDI_Green.pdf

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. doi:10.1108/10650750710720757

Guy, L., Kunszt, P., Laure, E., Stockinger, H., & Stockinger K. (2002). *Replica management in data grids. Global Grid Forum 5.* Edinburgh, Scotland. Retrieved from http://www.nesc.ac.uk/talks/ggf5_hpdc11/230702/rep_management_dg.pdf

Hacker, T. J., & Wheeler, B. C. (2007). Making research cyberinfrastructure a strategic choice. *Educause Quarterly*, *30*(1), 21-29. Retrieved from http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/MakingResearchCyberinfrastruct/157439

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 from http://www.dlib.org/dlib/march09/hank/03hank.html

Hedstrom, M. (1997). Digital preservation: A time bomb for digital libraries. *Computers and the Humanities*, *31*(3), 189-202. doi:10.1023/A:1000676723815

Hedstrom, M., & Montgomery, S. (1998). *Digital preservation needs and requirements in RLG member institutions: A study commissioned by the Research Libraries Group.* Mountain View, CA: Research Libraries Group. Retrieved from http://www.oclc.org/research/activities/past/rlg/digpresneeds/digpres.pdf

Hedstrom, M., Dawes, S., Fleischhauer, C., Gray, J., Lynch, C., McCrary, V., Moore, R., et al. (2003). *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation*. Washington DC: The National Science Foundation and The Library of Congress. Retrieved from http://www.digitalpreservation.gov/library/resources/pubs/docs/about_time2003.pdf

Helliwell, J. R., Strickland, P. R., & McMahon, B. (2006). *Position paper on IUCr response to the Global Information Commons for Science Initiative*. Chester, England. Retrieved from http://www.iucr.org/iucr/gicsi/positionpaper

Henty, M., Weaver, B., Bradbury, S., & Porter, S. (2008). *Investigating data management practices in Australian universities*. Canberra, Australia. Retrieved from http://hdl.handle.net/1885/47627

Hey, T. (2010). *Data-intensive scientific discovery: The fourth paradigm.* Bloomington, IN: Digital Science Center, Pervasive Technology Institute, Indiana University. Retrieved from http://pti.iu.edu/event/data-intensive-scientific-discovery-fourth-paradigm

Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G. C. Fox, & A. J. G. Hey (Eds.), *Grid Computing: Making the Global Infrastructure a Reality* (pp. 809-824). Chitchester, England: John Wiley & Sons, Ltd. Retrieved from http://eprints.ecs.soton.ac.uk/7648/1/The_Data_Deluge.pdf

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, *1*(3), 134-140. Retrieved from www.ijdc.net/index.php/ijdc/article/view/69/48

Holmes, V. P., Johnson, W. R., & Miller, D. J. (2004). Integrating metadata tools with the data services archive to provide Web-based management of large-scale scientific simulation data. *Proceedings of the 37th Annual Simulation Symposium (ANSS '04).* Arlington, VA, April 18-22: IEEE Computer Society. Retrieved from http://doi.ieeecomputersociety.org/10.1109/SIMSYM.2004.1299467

Holzner, A., Igo-Kemenes, P., & Mele, S. (2009). *Data preservation, reuse and (open) access in high-energy physics. CERN-OPEN-2008-028* (Vol. CERN-OPEN-). Geneva, Switzerland: CERN. Retrieved from http://cdsweb.cern.ch/record/1152295/files/CERN-OPEN-2008-028.pdf?version=1

Housewright, R., & Schonfeld, R. (2008). *Ithaka's 2006 studies of key stakeholders in the digital transformation in higher education*. New York, NY. Retrieved from http://reserved.serialssolutions.com/downloads/Ithaka-2006-Studies.pdf

Humphrey, C. (2006). *e-Science and the life cycle of research*. University of Alberta, Canada. Retrieved from datalib.library.ualberta.ca/~humphrey/**lifecycle**-science060308.doc

Hunter, J., & Choudhury, S. (2004). A semi-automated digital preservation system based on semantic web services. *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '04)* (pp. 269-278). Tucson, AZ, June 7-11: ACM & IEEE. doi:10.1145/996350.996415

Inmon, W. H., Strauss, D., & Neushloss, G. (2008). *DW2.0: The architecture for the next generation of data warehousing*. Burlington, MA: Morgan Kaufmann.

Interagency Working Group on Digital Data. (2009). *Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council*. Washington, D.C.: National Science and Technology Council. Retrieved from http://www.nitrd.gov/about/harnessing_power_web.pdf

Ives, Z. G., Halevy, A. Y., Mork, P., & Tatarinov, I. (2004). Piazza: Meditation and integration infrastructure for semantic web data. *Journal of Web Semantics*, *1*(2), 155 -175. doi:10.1016/j.websem.2003.11.003

Iwata, S. (2008). Editor's Note: Scientific "agenda" of data science. *Data Science Journal*, *7*(0), 54–56. Retrieved from http://www.jstage.jst.go.jp/article/dsj/7/0/7_54/_article

Jaiswal, A. R., Giles, C. L., Mitra, P., & Wang, J. Z. (2006). An architecture for creating collaborative semantically capable scientific data sharing infrastructures. In A. Bonifati & I. Fundulaki (Eds.), *Proceedings of the 8th Annual ACM International Workshop on Web Information and Data Management (WIDM 2006)* (pp. 75-82). Arlington, VA, November 10: ACM. Retrieved from http://doi.acm.org/10.1145/1183550.1183566

Jirotka, M., Procter, R., Rodden, T., & Bowker, G. (2006). Special Issue: Collaboration in e-research. *Computer Supported Cooperative Work (CSCW)*, *15*(4), 251-255. Retrieved from http://dx.doi.org/10.1007/s10606-006-9028-x

Jones, S., Ball, A., & Ekmekcioglu, C. (2008). The data audit framework: A first step in the data management challenge. *The International Journal of Digital Curation*, *2*(3), 112- 120. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/91/62

Kandasamy, K., Keerthikumar, S., Goel, R., Mathivanan, S., Patankar, N., Shafreen, B., Renuse, S., et al. (2009). Human Proteinpedia: A unified discovery resource for proteomics research. *Nucleic Acids Research*, *39*(Database Issue), D773–D781. Retrieved from http://nar.oxfordjournals.org/cgi/content/abstract/37/suppl_1/D773

Karasti, H., & Baker, K. S. (2008). Digital data practices and the long term ecological research program growing global. *The International Journal of Digital Curation*, *2*(3), 42 - 58. Retrieved from http://ijdc.net/index.php/ijdc/article/view/86

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructure in the Long Term Ecological Research (LTER) network. *Computer Supported Cooperative Work (CSCW)*, *15*(4), 321-358. Springer. doi:10.1007/s10606-006-9023-2

Kenney, A. R., McGovern, N. Y., Botticelli, P., Entlich, R., Lagoze, C., & Payette, S. (2002). Preservation risk management for web resources. *D-Lib Magazine*, *8*(1). Retrieved from http://www.dlib.org/dlib/january02/kenney/01kenney.html

Key Perspectives Ltd. (2010). *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study* (p. 31). Edinburgh, Scotland: Digital Curation Centre. Retrieved from http://hdl.handle.net/1842/3364

Kowalczyk, S. T. (2008). Digital preservation by design. In M. S. Raisinghani (Ed.), *Handbook of Research on Global Information Technology: Management in the Digital Economy* (pp. 405-431). Hershey, PA: Information Science Reference/IGI Global.

Kowalczyk, S. T., & Shankar, K. (2011). Data sharing in the sciences. In B. Cronin (Ed.), *Annual Review of Information Science and Technology. Volume 45.* (Vol. 45, pp. 247 - 294). Medford, NJ: Information Today, Inc.

LaRowe, G., Ambre, S., Burgoon, J., Ke, W., & Börner, K. (2009). The Scholarly Database and its utility for scientometrics research . *Scientometrics*, *79*(2), 219 - 234. Retrieved from http://www.akademiai.com/conten/W6056671M282N266

Lavoie, B., & Dempsey, L. (2004). Thirteen ways of looking at...digital preservation. *D-Lib Magazine*, *10*(7/8). Retrieved from http://www.dlib.org/dlib/july04/lavoie/07lavoie.html

Lercher, A. (2010). Efficiency of scientific communication: A survey of world science. *Journal of the American Society for Information Science and Technology*, *61*(10), 2049 - 2060. Retrieved from http://dx.doi.org/10.1002/asi.21384

Lesk, M. (2008). Recycling information: Science through data mining. *The International Journal of Digital Curation*, *3*(1), 154 - 157. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/71/50

Levitin, A. V., & Redman, T. C. (1993). A model of the data (life) cycles with application to quality. *Information and Software Technology*, *35*(4), 217-223. Retrieved from http://www.sciencedirect.com/science/article/B6V0B-48TDBN4-49/2/807c587d17d25d34b67fadefe3da14c2

Library of Congress. (2011). Preserving our digital heritage: The national digital information infrastructure and preservation program 2010 report. Washington, DC: A Collaborative Initiative of the Library of Congress. Retrieved from http://www.digitalpreservation.gov/library/resources/pubs/docs/NDIIPP2010Report_Post.pdf

Liu, L., & Chi, L. (2002). Evolutionary data quality: A theory-specific view. *Proceedings of the 7th International Conference on Information Quality, (MIT IQ Conference),* 292-304.

Long, D. E., Mantey, P. E., Wittenbrink, C. M., Haining, T. R., & Montague, B. R. (1995). REINAS: The Real-time Environmental Information Network and Analysis System. *40th IEEE Computer Society International Conference (COMPCON '95)*, 482. San Francisco, CA, March 05-09: IEEE Computer Society. Retrieved from http://www.computer.org/portal/web/csdl/doi/10.1109/CMPCON.1995.512426

Lor, P. J. & Snyman, M. M. M. (2005). Preservation of electronic documents in the private sector: Business imperative and heritage responsibility. *South African Journal of Information Management*, *7*(1). Retrieved from http://sajim.co.za/index.php/SAJIM/article/viewFile/253/244

Lord, P. & Macdonald, A. (2003). *e-Science curation report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. London, UK: Joint Information Systems Committee (JISC) Committee for the Support of Research. Retrieved from http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). *From data deluge to data curation*. (S. J. Cox, Ed.)*e-Science All Hands Meeting 2004* (pp. 371-375). Nottingham, England. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.7425&rep=rep1&type=pdf

Losh, L. (2010). Time will tell, but epistemology won't: Part II. *Virutalpolitik*. Retrieved from http://virtualpolitik.blogspot.com/2010/05/time-will-tell-but-epistemology-wont_15.html

Loshin, D. (2009). *Master data management*. Burlington, MA: Morgan Kaufmann Publishers/Elsevier.

Lynch, C. (2008). Big data: How do your data grow? *Nature*, *455*(7209), 28–29. Nature Publishing Group. Retrieved from http://www.nature.com/nature/journal/v455/n7209/full/455028a.html

Lyon, L. (2007). *Dealing with data: Roles, rights, responsibilities and relationships. Consultancy report.* Bath, UK: UKOLN and Joint Information Systems Committee (JISC) Committee for the Support of Research. Retrieved from http://ie-repository.jisc.ac.uk/171/1/13ealing_with_data_report-final.pdf

MacBean, N. (2008, November 10). Fridge-sized tape recorder could crack lunar mysteries. *Australian Broadcast Corporation News*. Sydney, Australia. Retrieved from http://www.abc.net.au/news/stories/2008/11/10/2415393.htm

Macey, R. (2006, August 5). One giant blunder for mankind: How NASA lost moon pictures. *The Sydney Morning Herald*. Sydney, Australia. Retrieved from

http://www.smh.com.au/news/national/one-giant-blunder-for-mankind-how-nasa-lost-moon-pictures/2006/08/04/1154198328978.html

Marcus, C., Ball, S., Delserone, L., Hribar, A., & Loftus, W. (2007). Understanding research behaviors, information resources, and service needs of scientists and graduate students: A study by the University of Minnesota Libraries. University of Minnesota Libraries. Retrieved from http://purl.umn.edu/5546

Martinez, L. (2009). *The Data Documentation Initiative (DDI) and institutional repositories* (pp. 1-21). Edinburgh, Scotland: JISC and EDINA, Edinburgh University. Retrieved from http://www.disc-uk.org/docs/DDI_and_IRs.pdf

Mathieu, R., & Khalil, O. (1998). Data quality in the database systems course. *Data Quality Journal*, *4*(1). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.3334&rep=rep1&type=pdf

Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological Informatics*, *1*(1), 3 - 7. doi:10.1016/j.ecoinf.2005.08.004

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

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: SAGE Publications.

Moore, R. (2008). Towards a theory of digital preservation. *The International Journal of Digital Curation*, *1*(3), 63 - 75. Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/63/42

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

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

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

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

Myers, J. D., Pancerella, C., Lansing, C., Schuchardt, K. L., & Didier, B. (2003). Multi-scale science: Supporting emerging practice with semantically derived provenance. In N. Ashish, M. Egenhofer, & C. Goble (Eds.), *Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*. Sanibel Island, FL, October 20. Retrieved from http://collaboratory.emsl.pnl.gov/resources/publications/papers/SCISW2003.Myersetal.pdf

Myers, J. D., 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. Retrieved from http://dx.doi.org/10.1007/s10586-005-4092-4

National Aeronautics and Space Administration. (1986). *Earth observing system. Data and information system. Volume 2A: Report of the EOS Data Panel. NASA Technical Memorandum 87777.* Washington DC: NASA Technical Memorandum Document ID 19860021622. Retrieved from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19860021622_1986021622.pdf

National Center for Biotechnology Information. (2008). *PubMed data provider documentation:* NLM Standard Publisher Data Format. Retrieved July 28, 2011, from http://www.ncbi.nlm.nih.gov/entrez/query/static/spec.html

National Information Standards Organization. (2008). *A framework of guidance for building good digital collections* (Vol. 2010). Baltimore, MD: National Information Standards Organization. Retrieved from http://framework.niso.org/node/5

National Science Board. (2005). *Long-lived digital data collections Enabling research and education in the 21st century*. Arlington, VA: National Science Board Committee on Programs and Plans, NSB-05-40. Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_1.pdf

National Science Foundation. (1998). *Digital libraries initiative – phase 2.* (Vol. 2010). Washington, D.C. Retrieved from http://www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm

National Science Foundation. (2010). *NSF organization list*. Retrieved from http://www.nsf.gov/staff/orglist.jsp

Niederman, F., Mathieu, R., & Morley, R. (2007). Examining RFID applications in supply chain management. *Communications of the ACM*, *50*(7), 92-101. Retrieved from http://doi.acm.org/10.1145/1272516.1272520

Niu, X., Hemminger, B. M., Lown, C., Adams, S., Brown, C., Level, A., McLure, M., et al. (2010). National study of information seeking behavior of academic researchers in the United States. *Journal of the American Society for Information Science and Technology*, *61*(5), 869 - 890. Retrieved from http://dx.doi.org/10.1002/asi.21307

Otto, B., Wende, K., Schmidt, A., & Osl, P. (2007). Towards a framework for corporate data quality management. *18th Australasian Conference on Information Systems. The University*

of Southern Queensland, Toowoomba, Australia (Vol. 109, pp. 916–926). Toowoomba, Australia, December 5-7: Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.4951&rep=rep1&type=pdf.

Pearson, D. (2007). AONS II: Continuing the trend towards preservation software "Nirvana." *International Conference on Preservation of Digital Objects (iPres2007),* Beijing, China, October 11-12. Retrieved from http://en.scientificcommons.org/32972083

Pielke, R. A. (2002). *Mesoscale meteorological modeling*. San Diego: Academic Press.

Podsakoff, P. M. (2005). *Item-construct Matrix.* Seminar in Behavioral Research Methods, Fall 2005. Indiana University, Kelley School of Business.

Pritchard, S. M., Anand, S., & Carver, L. (2005). *Informatics and knowledge management for faculty research data (ID: ERB0502)*. Boulder, CO: EDUCAUSE Center for Applied Research: Research Bulletin, Vol. 2. Retrieved from http://net.educause.edu/ir/library/pdf/ERB0502.pdf

Pryor, G. (2007). Project StORe: Making the connections for research. *OCLC Systems & Services*, *23*(1), 70-78. doi:10.1108/10650750710720775

Pryor, G., & Donnelly, M. (2009). Skilling up to do data: Whose role, whose responsibility, whose career? *International Journal of Digital Curation*, *4*(2), 158-170. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/126

Rabinovici-Cohen, S., Factor, M. E., Naor, D., Ramati, L., Reshef, P., Ronen, S., Satran, J., et al. (2008). Preservation dataStores: New storage paradigm for preservation. *IBM Journal of Research and Development*, *52*(4/5), 1 - 11. Retrieved from http://www.research.ibm.com/haifa/projects/storage/datastores/papers/rabinovici.pdf

Rajasekar, A. K., & Moore, R W. (2001). Data and metadata collections for scientific applications. In B. Hertzberger, A. Hoekstra, & R. Williams (Eds.), *Proceedings of the 9th International Conference on High-Performance Computing and Networking (HPCN Europe 2001)* (pp. 72-80). Amsterdam, The Netherlands, June 25-27: Springer Berlin / Heidelberg. Retrieved from http://www.springerlink.com/content/2dpf93lxdqjrrb25

Rajasekar, A., Marciano, R., & Moore, R. (1999). Collection-based persistent archives. In A. Kerr, B. Kobler, & R. Moore (Eds.), *Proceedings of the 16th IEEE Mass Storage Systems Symposium and 7th NASA Goddard Conference on Mass Storage Systems and Technologies*. San Diego, CA, March 15-18: IEEE Computer Society. Retrieved from http://storageconference.org/1999/papers/17rajase.pdf

Reed, M. (2010). Data deluge. Retrieved from http://www.datadeluge.com/

Research Libraries Group & OCLC. (2002). *Trusted digital repositories: Attributes and responsibilities: An RLG-OCLC Report*. Mountain View, CA: Research Libraries Group

and OCLC. Retrieved from http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf

Rice, R. (2007). DISC-UK DISC-UK DataShare Projects: Building exemplars for institutional data repositories in the UK. *IASSIST Quarterly*, *2007*(Fall/Winter), 21 - 27. Retrieved from http://www.iassistdata.org/content/disc-uk-datashare-project-building-exemplars-institutional-data-repositories-uk

Riley, B. (2006). Trusting data's quality: Database publication presents unique challenges for the peer reviewer. *Nature*. doi:doi:10. 1038/nature04993

Rosenthal, D. S. H., Robertson, T. S., Lipkis, T., Reich, V., & Morabito, S. (2005). Requirements for digital preservation systems: A bottom-up approach. *D-Lib Magazine*, *11*(11). Retrieved from www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

Rosenthal, D. S., Roussopoulos, M., Giuli, T., Maniatis, P., & Baker, M. (2004). Using hard disks for digital preservation. *IS&T Archiving Conference Final Program and Proceedings (Archiving 2004)*, 249-253. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9782&rep=rep1&type=pdf

Ross, S. (2007). Digital preservation, archival science and methodological foundations for digital libraries. *Proceedings of the 11th European Conference on Digital Libraries (ECDL), Budapest (17 September 2007).* Budapest, Hungary: Springer. Retrieved from http://www.ecdl2007.org/Keynote_ECDL2007_SROSS.pdf

Rumsey, A. S. (2010). *Sustainable economics for a digital planet: Ensuring long-term access to digital information. Final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access.* Washington, DC: National Science Foundation (NSF Award No. OCI 0737721). Retrieved from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Rusbridge, C. (2007). Create, curate, re-use: The expanding life course of digital research data. *EDUCAUSE Australasia 2007*. Melbourne, Victoria, Australia: EDUCAUSE. Retrieved from http://hdl.handle.net/1842/1731

Ryssevik, J. (2001). *The Data Documentation Initiative (DDI) metadata specification*. Ann Arbor, MI: Data Documentation Alliance. Retrieved from http://www.ddialliance.org/sites/default/files/ryssevik_0.pdf

Sarker, S., Lau, F., & Sahay, S. (2001). Using an adapted grounded theory approach for inductive theory building about virtual team development. *ACM SIGMIS Database*, *32*(1), 38-56. doi: http://doi.acm.org/10.1145/506740.506745

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

Simmhan, Y., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record*, *34*(3), 31-36. ACM Press. Retrieved from citeulike-article-id:678653

Soylu, A., De Causmaecker, P., & Desmet, P. (2009). Context and adaptivity in pervasive computing environments: Links with software engineering and ontological engineering. *Journal of Software*, *4*(9 (2009)), 992-1013. doi:10.4304/jsw.4.9.992-1013

Stanescu, A. (2005). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *OCLC Systems & Services*, *21*(1), 61-81. Retrieved from http://www.emeraldinsight.com/Insight/ViewContentServlet?contentType=Article&Filename=/published/emeraldfulltextarticle/pdf/1640210110.pdf

Stanford University Libraries. (2005). *ISS – Library of Congress archive ingest and handling test (AIHT): Final report of the Stanford Digital Repository* (p. 98). Washington, D.C. Retrieved from http://www.digitalpreservation.gov/partners/aiht/high/aiht-stanford-final-report.pdf

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

Stern, P. N. (2007). On solid ground: Essential properties for growing grounded theory. In A. Bryant & K. Charmaz (Eds.), *The Sage handbook of grounded theory* (pp. 114 - 126). Thousand Oaks, CA.

Strauss, Anselm, & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques* (p. 270). Newbury Park, CA: Sage Publications.

Swan, A., & Brown, S. (2008). *To share or not to share: Publication and quality assurance of research data outputs: Main report*. London, UK: Research Information Network, Joint Information Systems Committee (JISC) Committee for the Support of Research, and the National Environment Research Council UK. Retrieved from http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-report.pdf.

Thomas, W., Gregory, A., & Piazza, T. (2005). Inside view of DDI Version 3.0: Structural Reform Group report. *International Association for Social Science Information Services and Technology (IASSIST)*. Edinburgh, UK: International Association for Social Science Information Services and Technology (IASSIST). Retrieved from http://www.iassistdata.org/conferences/conference-2005-presentations

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

Urquhart, C. (2007). The evolving nature of grounded theory method: The case of the information systems discipline. In A. Bryant & K. Charmaz (Eds.), *The Sage Handbook of Grounded Theory* (pp. 339-359). Thousand Oaks, CA: Sage Publications, Ltd.

Urquhart, C., & Fernandez, W. (2006). Grounded theory method: The researcher as blank slate and other myths. *Proceedings of International Conference on Information Systems 2006*. Milwaukee, WI, December 10-13. Retrieved from http://aisel.aisnet.org/icis2006/31

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a standard for the social sciences. *The International Journal of Digital Curation*, *3*(1), 107-113. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/66/45

Venkatraman, N., & Grant, J. H. (1986). Construct measurement in organizational strategy research: A critique and proposal. *Academy of Management Review*, *11*, 71 – 87.

Venugopal, S., Buyya, R., & Ramamohanarao, K. (2006). A taxonomy of data grids for distributed data sharing, management, and processing. *ACM Computing Surveys (CSUR)*, *38*(1). doi: http://doi.acm.org/10.1145/1132952.1132955

Voss, C., Tsikriktsis, N., & Frohlich, M. (2002). Case research in operations management. *International Journal of Operations & Production Management*, *22*(2), 195 - 219. Retrieved from http://dx.doi.org/10.1108/01443570210414329

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. Retrieved from citeulike-article-id:6338125

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623 - 640. doi:10.1109/69.404034

Watry, P. (2007). Digital preservation theory and application: Transcontinental persistent archives testbed activity. *The International Journal of Digital Curation*, *2*(2), 41 - 68. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/43/28

Weber, R. (2003). Editor's comments: Theoretically speaking. *MIS Quarterly*, *27*(3), iii-xii. Retrieved from misq.org/misq/downloads/download/editorial/32/

Weinberg, A. M. (1961). Impact of large-scale science on the United States. *Science*, *134*(3473), 161-164. Retrieved from http://www.jstor.org/stable/1708292

van Westrienen, G., & Lynch, C. A. (2005). Academic institutional repositories: Deployment status in 13 nations as of mid-2005. *D-lib Magazine*, *11*(9). Retrieved from http://www.dlib.org/dlib/september05/westrienen/09westrienen.html

Whyte, A., Job, D., Giles, S., & Lawrie, S. (2008). Meeting curation challenges in a neuroimaging group. *The International Journal of Digital Curation*, *3*(1), 171 - 181. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/74/53

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *The International Journal of Digital Curation*, *3*(4), 93 - 103. Retrieved from http://ijdc.net/index.php/ijdc/article/view/137/165

Appendix A. Approved IRB Forms

INDIANA UNIVERSITY
OFFICE OF RESEARCH ADMINISTRATION

To: Stacy T. Kowalczyk
SLIS

From: IUB Human Subjects Office
Office of Research Administration – Indiana University

Date: November 22, 2010

RE: **EXEMPTION GRANTED – NEW PROTOCOL**
Protocol Title: Digital Data Research Environments Study
Protocol #: 1010002804
Sponsor:

Your study named above was accepted on November 21, 2010 as meeting the criteria of exempt research as described in the Federal regulations at 45 CFR 46.101(b), paragraph(s) 2. This approval does not replace any departmental or other approvals that may be required.

As the principal investigator (or faculty sponsor in the case of a student protocol) of this study, you assume the following responsibilities:

- Changes to Study: Any proposed changes to the research must be approved by the IRB prior to implementation. To request approval, please complete an Amendment form and submit it, along with any revised study documents to iub_hsc@indiana.edu. Only after approval has been granted by the IRB can these changes be implemented.

- Completion: Although a continuing review is not required for an exempt study, you are required to notify the IRB when this project is completed. In some cases, you will receive a request for current project status from our office. If we are unsuccessful in our attempts to confirm the status of the project, we will consider the project closed. It is your responsibility to inform us of any changes to your contact information to ensure our records are kept current.

Per federal regulations, there is no requirement for the use of an informed consent document or study information sheet for exempt research, although one may be used if it is felt to be appropriate for the research being conducted. As such, these documents do not include an IRB-approval stamp. Please note, however, that if a study information sheet and/or informed consent document is to be used, you should use unstamped accepted versions. **Please note that your study has been accepted with the use of a study information sheet / informed consent document.**

You should retain a copy of this letter and any associated approved study documents in your records. Please refer to the protocol title and number in future correspondence with our office. You may contact our office at (812) 855-3067 or by e-mail at iub_hsc@indiana.edu if you have questions or need further assistance.

Thank you.

INDIANA UNIVERSITY BLOOMINGTON INSTITUTIONAL REVIEW BOARD (IRB) REVIEW

EXEMPT RESEARCH CHECKLIST

IRB Study #: 1010002804
(IRB Office will assign)

DIRECTIONS: This form is to be neatly typed and submitted to the IRB only when the investigator is contemplating the initiation of a research project which, in the investigator's judgment, is exempt from full IRB review. The IRB will then determine whether the activity is covered by these regulations.

Research activities are exempt from regulations for the protection of human research subjects when they are considered minimal risk (the probability or magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests (as defined by 45 CFR 46.102(i)) and the ONLY involvement of human subjects falls within one or more of the exempt categories listed below.

The exempt categories outlined below do not apply to research involving prisoners or research involving a test article regulated by the FDA, unless the research meets the criteria for exemption described in 45 CFR 46.101(b)(6) and 21 CFR 56.104(d).

The exempt categories outlined below are based solely on methods of research, and do not take the level of risk into consideration. Although most exempt research requires no further oversight to be conducted ethically, some exempt research raises ethical concerns or requires measures to protect participants. As such, the IRB will not consider any research exempt that does not fulfill ethical principles reflected in the Belmont Report. These basic ethical principles are:

1. Respect for Persons (Autonomy) – individuals should be treated as autonomous agents and persons with diminished autonomy are entitled to protection.
2. Beneficence – Human subjects should not be harmed and the research should maximize possible benefits and minimize possible harms.
3. Justice – the benefits and risks of research must be distributed fairly.

Research that otherwise would be exempt by federal regulations that raises ethical concerns or requires measures to protect subjects may be denied and/or moved to a higher level of review (i.e. expedited or full IRB review).

Check the appropriate category(ies) that applies to your research project:

☐	1. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as (i) research on regular and special educational instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods. [45CFR46.101(b)(1)]
☒	2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless all of the following are true: (i) information obtained is recorded in such a manner that the human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, insurability, or reputation. [45CFR46.101(b)(2)] **NOTE: If the research involves children as participants, the research must be limited to educational tests (cognitive, diagnostic, aptitude, achievement) and observation of public behavior when the investigator(s) do not participate in the activities being observed. Research involving children that uses survey procedures, interview procedures, or observation of public behavior when the investigator(s) participate in the activities being observed cannot be granted an exemption.**
☐	3. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under category 2 above, if either:

	(i) the human subjects are elected or appointed public officials or candidates for public office; or (ii) federal statute(s) require(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter. [45CFR46.101(b)(3)]

If any of the above categories have been selected, answer the following:

Will you be audio or video recording?
☒ No
☐ Yes. Explain how it will be assured that the identity of the subjects and/or link to the information obtained or the information recorded about the subjects does not place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, insurability, or reputation:

☐	4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. [45CFR46.101(b)(4)] **To qualify for this exemption, data, documents, records, or specimens must exist at the time the research is proposed and not prospectively collected.** Provide a list of all data points that will be collected below or attach a data collection sheet.
☐	5. Research and demonstration projects which are conducted by or subject to the approval of Department or Agency heads, and which are designed to study, evaluate, or otherwise examine: (i) public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs. [45CFR46.101(b)(5)]. The program under study must deliver a public benefit (for example, financial or medical benefits as provided under the Social Security Act) or service (for example, social, supportive, or nutrition services as provided under the Older Americans Act). The research or demonstration project must be conducted pursuant to specific federal statutory authority, must have no statutory requirement that an IRB review the project, and must not involve significant physical invasions or intrusions upon the privacy of the subjects. This exemption is for projects conducted by or subject to approval of Federal agencies and requires authorization or concurrence by the funding agency.
☐	6. Taste and food quality evaluation and consumer acceptance studies, (i) if wholesome foods without additives are consumed; or (ii) if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural, chemical, or environmental contaminant at or below the level found to be safe, by the Food and Drug Administration or approved by the Environmental Protection Agency or the Food Safety and Inspection Service of the U.S. Department of Agriculture. [45CFR46.101(b)(6) and 21 CFR 56.104(d)]

INDIANA UNIVERSITY – BLOOMINGTON INSTITUTIONAL REVIEW BOARD (IRB) REVIEW

EXEMPT RESEARCH CHECKLIST

IRB Study #: 1010002804
(IRB Office will assign)

SECTION I: INVESTIGATOR INFORMATION

Principal Investigator: Kowalczyk, Stacy T. **Department:** SLIS
(Last, First, Middle Initial)
Building/Room No.: Innovation Center Room 130L-G Phone: (812) 856-2146 E-Mail: skowalcz@indiana.edu

Faculty Sponsor: Katy Börner **Department:** SLIS
(Last, First, Middle Initial)
Building/Room No.: LI 021 Phone: (812) 855-3256 E-Mail: katy@indiana.edu

Project Duration: Start Date: **October 15, 2010** End Date: **October 15, 2011**

Project Title: **Digital Data Research Environments Study**

Sponsor/Funding Agency: N/A

SECTION II: PERFORMANCE SITE

☒ Indiana University Bloomington Campus; state location(s):

☐ Other Indiana University Campus: state location(s):

☐ Anthropology
☐ Bloomington Hospital
☐ Bradford Woods
☐ School of Business
☐ Economics
☐ School of Education
☐ French and Italian
☐ Gender Studies
☐ Health Center
☐ Health, Phys Ed & Rec (HPER)
☐ IN Institute on Disability & Communication
☐ Informatics
☐ School of Journalism
☐ The Kinsey Institute
☐ Library General
☒ School of Library & Info Science
☐ MCCSC (Monroe School District)
☐ School of Music
☐ Nursing
☐ Optometry
☐ Population Institute for Research & Training
☐ Department of Psychological and Brain Sciences
☐ Second Language Studies
☐ Sociology
☐ Spanish & Portuguese
☐ Public & Environmental Affairs (SPEA)
☐ Speech and Hearing Sciences
☐ Center for Survey Research
☐ Telecommunications
☐ University Info Tech Services
☐ Center for Evaluation and Education Policy
☐ Central Eurasian Studies
☐ Communication and Culture
☐ Computer Science
☐ Criminal Justice
☐ Folklore and Ethno Musicology
☐ History
☐ Linguistics

☐ Other:

SECTION III: RESEARCH DESCRIPTION

1. Provide a brief description, in lay terms, of the purpose of the proposed project and the procedures to be used.

   Preserving scientific digital data, ensuring its continued access, has emerged as a major initiative for both funding agencies and academic institutions. Providing long-term access to digital data has a number of challenges. Digital data requires constant and perpetual maintenance. Technologies change; equipment ages; software is superseded. Digital data is not fixed and can easily be

changed, either intentionally or unintentionally. Much of the research in digital preservation has focused on repositories – systems to manage digital content, to collect and store sufficient technical metadata for preservation, and to manage and initiate preservation actions. This research stream presumes that the data has been either created in a Cyberinfrastructure environment or pushed into a preservation environment and does not address the antecedents to preservation. Yet these antecedents are crucial to the act of preservation. The antecedents to preservation – data management, contextual metadata, and access to preservation technologies – can also be barriers preventing preservation.

This dissertation will develop a theoretical model of the e-Science data environment describing the antecedents to preservation. The data environment is the set of technologies and socio-technical universe in which scientists create, use and store their data. The data environment is a complex interaction of content, formats, context, quality control, data collections, and the technical infrastructure of the lab's home institution. This dissertation proposes a new survey to test, generalize, and enhance this model. The questions that this research will address are:

1. How do scientists perceive the need to preserve their data?
2. How do scientists manage their data?
3. In what environments does the data exist?
4. How do scientists create data and control its quality?

This dissertation proposes to conduct a web-based survey targeted to the Principle Investigators (PIs) of National Science Foundation grant awards. The survey in its web format is attached as Appendix A. Questions can be skipped if the participant chooses not to answer. The survey should take approximately 20 minutes to complete. The invitation to participate will be sent via email. The study information sheet (attached as Appendix B) will be sent as an attachment to the email. The participants will be asked to read the sheet before initiating the survey. The information sheet will also be available from the web survey. The survey can be accessed via this URL: https://iucsr.qualtrics.com/SE/?SID=SV_24dGrJFpNxDw6Uc

The data will be analyzed using traditional statistical methods such as contingency tables, factor analysis, and linear regression modeling.

No personal or indentifying information will be collected about the participants. No persons under 18 will be invited to participate.

It is expected that this survey will take approximately 20 minutes to complete.

a. Please state the eligibility (inclusion/exclusion criteria).

   Practicing researchers will be invited to participate in this study. The sample will be seeded with awardees (Principle Investigators) of National Science Foundation grants. Approximate 8,400 recent awardees will be invited to participate via email using the Scholarly Database developed and maintained by Katy Börner' lab. The data in the Scholarly Database is public information provided by the National Science Foundation. The database has the PI's name, email address, and the NSF directorate that awarded the funding. The data is also available on the NSF's public website (http://www.nsf.gov/awardsearch/). The PIs will be encouraged to forward the invitation to others in their labs and research community.

b. Will subjects be compensated for participation?

   No

ONLY COMPLETE 2-4 BELOW IF YOU SELECTED CATEGORY 1, 2, 3, 5, OR 6 ON THE EXEMPT RESEARCH CHECKLIST.

2. Provide the process by which individuals will be recruited.

   A set of approximately 8,500 names and email addresses will be selected at random from each of the 7 NSF directorates using the Scholarly Database which is created and maintained by the Indiana University Cyberinfrastructure for Network Science Center. For each of the 7 NSF directorates, all of the names with email addresses will be pulled. A script will process the list to pull every 11th name until at least 1,200 names are pulled. In the unlikely situation that this process will not produce the desired 1,200 names per directorate, the script will be modified to additionally pull every 15th name until the 1,200 is reached.

All of the contact information (including email addresses) to be used in this study was harvested from the National Science Foundation publically available website. This data is in the public record and available for use by the public.

An email request for participation will be sent to each person selected. The email message is attached as Appendix C. As explained in detail in Section 3 below, up to 3 emails could be sent to each individual.

a. Explain how it will be ensured that recruitment or selection will not unfairly target a particular population or will target the population that will benefit from the project/research.

This study is to understand the data management practices of active, federally funded researchers. Therefore, the population targeted is active, federally funded researchers. This study will seed a snowball sample with the names and email addresses of recent recipients of National Science Foundations grants. These Principle Investigators will be asked to forward the participation request email to other researchers in their lab or research group and their research community.

3. Explain how it will be ensured that individuals will be treated with respect during interactions/observations with them. For those individuals with diminished autonomy (e.g. children, people with limited ability to make decisions), explain how they will be protected.

This study will not have direct interaction with the participants. An initial email solicitation will be sent inviting people to click on a link to the web-based survey using the Qualtrix email management function as is the practice of the Center for Survey Research at Indiana University. A follow-up email will be sent to those who have not yet participated reminding them of the opportunity. The follow-up email message is attached in Appendix C.

The survey itself will not collect or retain email addresses. The Qualtrix email management function will track the responses via email address until the end of the survey when the data will be deleted. The responses will not be traceable to an email address.

Children and people with limited ability to make decisions will not be solicitied and are not expected to be included in this research.

a. Explain how individual privacy will be protected. For example, if interviewing, where will that be conducted?

This survey will only be administered via the web. The participants will take the survey in the environment which is most comfortable for them. Privacy is ensured by the choice of the participant.

b. Explain how individual confidentiality will be protected. For example, what kind of information will be recorded and how will that be protected?

This survey will collect no identifying information about participants. Initially, the data will be collected and stored on a secure survey service used by the Indiana University Center for Survey Research called Qualtrix (http://new.qualtrics.com/). Once the survey is closed, the data will be deleted from the Qualtrix servers and stored on a secured IU server. This data will be kept for 2 years and then deleted. Any information collected byt the Qualtrix email system to manage follow-up email requests will be deleted as soon as the survey is complete.

4. How will you help to minimize potential risks that individuals may be exposed to while participating in the research? Potentials risks may include psychological, social, legal, physical, etc.

The potential risks for participants are minimal. While it could be possible for participants to feel concern about the longevity of their data, their inability to answer the questions with specificity, or their inability to determine preservation risks, there is no expectation of any harm.

All questions asking for specifics will have moderating words such as "if possible..." or "if you are able..." indicating that the survey anticipates that this information may be difficult to convey.

SECTION IV: CO-INVESTIGATORS

A. Co-investigators: Provide the name and department of other individual(s) assisting with the study who 1) will be responsible for the design, conduct, or reporting of the study, 2) have access to subjects (i.e. will consent subjects, conduct parts of the study), 3) will be making independent decisions about the inclusion or exclusion of participants, or 4) have access to identifying and confidential information.

1. List individuals from affiliated institutions who are directly interacting or intervening with subjects:
   Name Department

   The individuals listed above are required to:

   1) **Pass the IU human subjects protection test, unless special circumstances apply. Please refer to http://www.iupui.edu/%7Eresgrad/Human%20Subjects/human-menu.htm for additional information.**
   2) **Provide the IRB with documentation of their agreement to participate in the research. This can be accomplished by having the individual provide his/her signature next to his/her name above or including a memo (or email) from the individual documenting agreement to participate in this specific protocol.**
   3) **Have a Conflict of Interest (COI) disclosure form on file with the COI Committee. Please refer to http://www.iupui.edu/~resgrad/spon/policiescontent.htm for additional information.**

2. List individuals from affiliated institutions who are **not** directly interacting or intervening with subjects:
   Name Department

B. **Collaborating Co-investigators**. List any co-investigators from nonaffiliated institutions for which the IU-Bloomington IRB is providing the review and approval for their role in the study.

   Note: For each nonaffiliated investigator, a nonaffiliated investigator agreement may be required. For additional guidance, refer to: http://www.iupui.edu/%7Eresgrad/spon/non-affiliated-pi.rtf. Nonaffiliated investigators who are directly interacting or intervening with subjects (including obtaining consent) must either pass the IU humans subjects protection test, be from a COGR institution, or provide documentation of passing the CITI or NCI protection of human subjects test.

   Name of Co-investigator Institution Role Procedures performed

---

Statement of Principal Investigator. I have personally reviewed this application and agree with its contents and am aware of my responsibility to provide supervision and guidance during its execution (in the case of a student project).

Principal Investigator Signature:_____ Date:_____

Faculty Sponsor Signature:_____ Date:_____

Note: As an alternative to providing original signatures on the form, the PI should simply e-mail the completed form to iub_hsc@indiana.edu. This e-mail serves as the PI's signature. For the faculty sponsor's signature, please forward an e-mail from the individual acknowledging his/her oversight responsibilities for the student research project. This will serve as the faculty sponsor's signature.

Appendix B. Study Participant Solicitation Email Text

First Email Solicitation

Dear Dr. NAME,

As part of my Ph.D. dissertation research (IRB study #_____), I am conducting a survey on the nature of digital research data. The purpose of this study is to uncover issues surrounding the data management practice of researchers, data quality, and the long-term retention of data.

As a recent awardee of a National Science Foundation grant, you are invited to participate in this study – a web-based survey that should take approximately 20 minutes to complete. Feel free to forward this message to others in your research lab or group and to your research community.

Your answers are strictly confidential. No report will identify any individual person, research unit, or academic institution. Please read information study sheet linked off the survey site before participating in the survey (http://tinyrul.com/24hhe4f). If you have any questions please feel free to contact me either through email (skowalcz@indiana.edu) or phone (812) 856-2146.

The online survey is available at https://iucsr.qualtrics.com/SE/?SID=SV_24dGrJFpNxDw6Uc

Thank you for your time,

Stacy Kowalczyk

Subsequent Follow Up Email Solicitation

Dear Dr. NAME,

I would like to invite you to participate in a study to understand the data management practice of researchers, issues of data quality and the long-term retention of data. This is a web-based survey that should take approximately 20 minutes to complete. This survey is part of Study #_____.

Feel free to forward this message to others in your research lab or group and to your research community.

Your answers are strictly confidential. No report will identify any individual person, research unit, or academic institution. Please read information study sheet linked off the survey site

before participating in the survey (http://tinyrul.com/24hhe4f). If you have any questions please feel free to contact me either through email (skowalcz@indiana.edu) or phone (812) 856-2146.

The online survey is available at https://iucsr.qualtrics.com/SE/?SID=SV_24dGrJFpNxDw6Uc

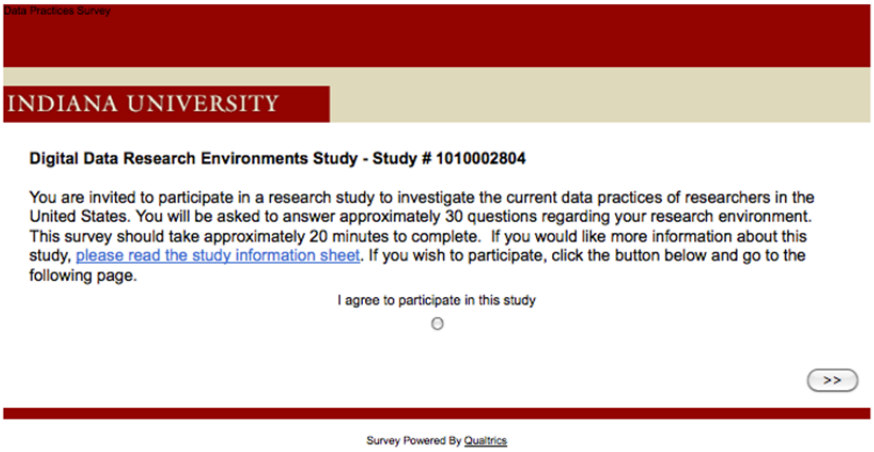
Thank you for your time,

Stacy Kowalczyk

Appendix C. Molecular Biology Description Example

cellular & mol biol
micro & molec biol
microbio & mol gen
mol biology
mol biophysics
mol, cell & devel biol
mol, micro & struct biol
mole biology
mole genet & microbiol
molec biol
molec biology
molec cell & develop biol
molec genetics & cell biology
molec genetics/microbiology
molec microbio& immunology
molec pharm and biol chem
molec, cell & dev biology
molec, cell & develop biology
molecular & cell biology
molecular & cellular biology
molecular and cell biology
molecular and cellular biol
molecular and cellular biology
molecular bio
molecular biol
molecular biolog
molecular biology
molecular biology & biochem
molecular biology & biochemistr
molecular biology & biophysics
molecular biology & genetics
molecular biology & microbio
molecular biology/microbiology
molecular biophysics & biochem
molecular biophysics and bioche
molecular genetics & cell bio
molecular genetics & cell biol
molecular genetics and cell bio
molecular genetics/cell bio
molecular microb & immunology
molecular microbiology
molecular microbiology & immuno
molecular physiology & biophysi
molecular, cellular & dev bio
molecularbio
molecularbiology

Appendix D. Survey Instrument



INDIANA UNIVERSITY

Which term below best describes your position?

- ○ Principal investigator
- ○ Researcher
- ○ Post Doc
- ○ Ph.D. Student/candidate
- ○ Masters Student
- ○ Other

What is your primary research institution?

What is your scientific domain?

(Please select the domain that most closely describes your research. These are based on NSF directorates)

How is your research funded?

- ☐ Exclusively grant funded
- ☐ Mostly grant funded with some ongoing funding from my institution
- ☐ Mostly funded by my institution with some grant funding
- ☐ Equally funded by grants and by my institution
- ☐ Exclusively funded by my institution
- ☐ Not sure

What is the size of your research group or lab?

- ○ I do not work in a group or a lab
- ○ 5 or more researchers
- ○ Less than 5 researchers

This survey will ask questions about your research process. Please think of a recent project, either completed or ongoing, which best exemplifies your research. Please use this project to answer questions that refers to "your project".

Please provide a brief description of this project:

What research methods did you use in your project? [check all they apply]

- ☐ Surveys
- ☐ Field studies
- ☐ Case studies
- ☐ Direct observation in experimental situations
- ☐ Analysis of instrument generated data
- ☐ Analysis of existing data sets
- ☐ Modeling and simulation
- ☐ Text or language analysis
- ☐ Other

<< >>

INDIANA UNIVERSITY

Do you worry about the longevity of your data?

- ○ Quite a lot
- ○ Somewhat
- ○ Not much
- ○ Not at all
- ○ Not sure
- ○ Other

The best term to describe your level of concern about preserving your data is

- ○ Very concerned
- ○ Moderately concerned
- ○ Slightly concerned
- ○ Not concerned at all
- ○ Not sure
- ○ Other

Do you have any contractual obligations (through grants or other agreements) to keep your data usable for a specific length of time?

- ○ Yes. Please indicate how long you need to keep the data usable.
- ○ No
- ○ Not sure
- ○ Other

Is it important to you to make your research data available to future generations of researchers?

- ○ Very important
- ○ Somewhat important
- ○ Not very important
- ○ Not important at all
- ○ Not sure
- ○ Other

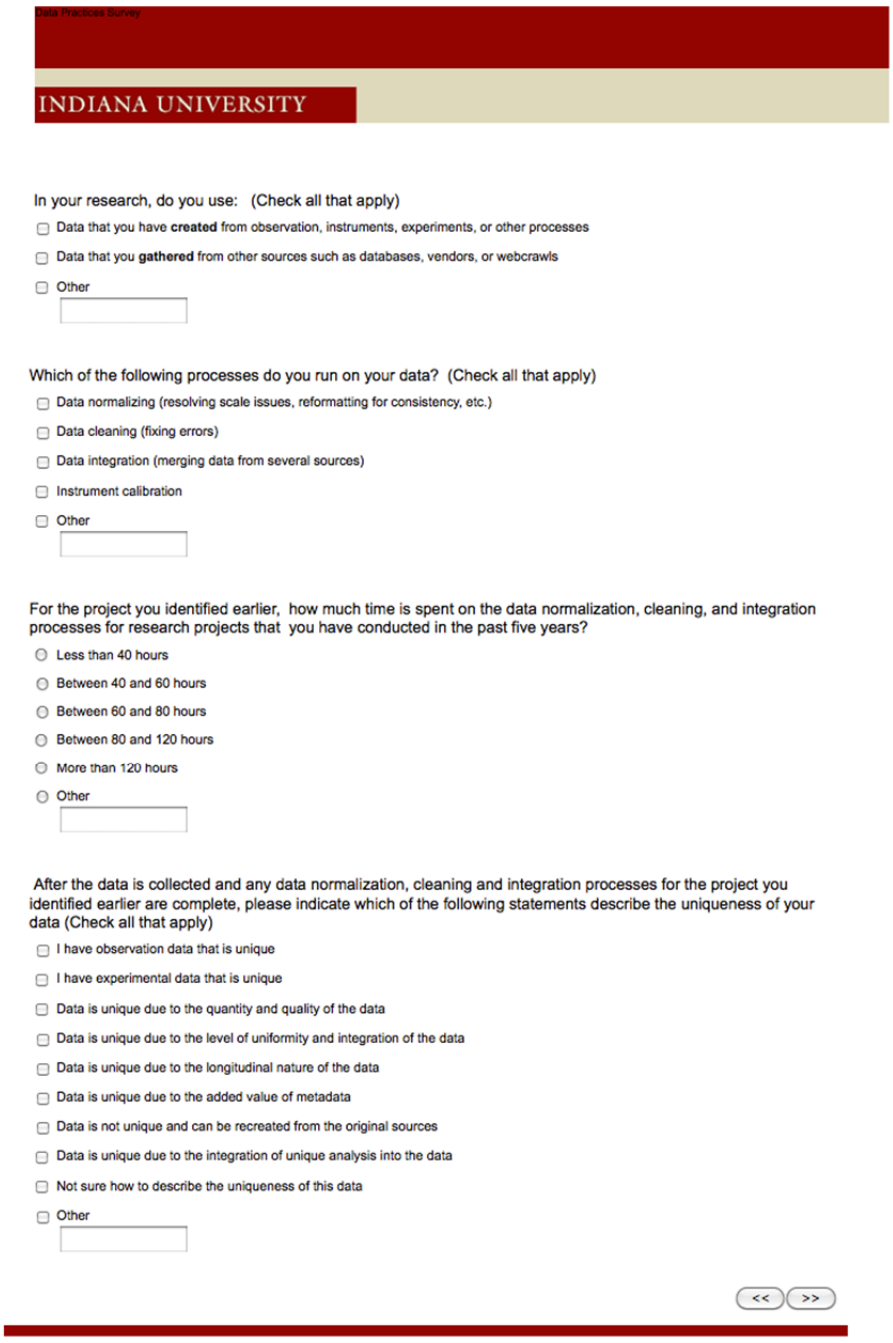
For your research data, how easy is it for you to identify the ***most important*** data to preserve?

- ○ Very Easy
- ○ Somewhat easy
- ○ Somewhat difficult
- ○ Very difficult
- ○ Not sure
- ○ Other

For your research data, how easy is it for you to identify the data that is ***most in need*** of preservation?

- ○ Very easy
- ○ Somewhat easy
- ○ Somewhat difficult
- ○ Very difficult
- ○ Not sure
- ○ Other

<< >>


Data Practices Survey
INDIANA UNIVERSITY
In your research, do you use: (Check all that apply)
Data that you have created from observation, instruments, experiments, or other processes
Data that you gathered from other sources such as databases, vendors, or webcrawls
Other
Which of the following processes do you run on your data? (Check all that apply)
Data normalizing (resolving scale issues, reformatting for consistency, etc.)
Data cleaning (fixing errors)
Data integration (merging data from several sources)
Instrument calibration
Other
For the project you identified earlier, how much time is spent on the data normalization, cleaning, and integration processes for research projects that you have conducted in the past five years?
Less than 40 hours
Between 40 and 60 hours
Between 60 and 80 hours
Between 80 and 120 hours
More than 120 hours
Other
After the data is collected and any data normalization, cleaning and integration processes for the project you identified earlier are complete, please indicate which of the following statements describe the uniqueness of your data (Check all that apply)
I have observation data that is unique
I have experimental data that is unique
Data is unique due to the quantity and quality of the data
Data is unique due to the level of uniformity and integration of the data
Data is unique due to the longitudinal nature of the data
Data is unique due to the added value of metadata
Data is not unique and can be recreated from the original sources
Data is unique due to the integration of unique analysis into the data
Not sure how to describe the uniqueness of this data
Other
<< >>

INDIANA UNIVERSITY

In the project you identified earlier, approximately how many times did you convert data from one format into another?

- ○ I do not convert data at all
- ○ Less than 3 times
- ○ Between 3 and 5 times
- ○ More than 5 times
- ○ Not sure
- ○ Other

For the project you identified earlier, which of the scenarios below would best describe the format conversion process?

- ○ I do not convert data at all
- ○ I convert data from a single source into a single standard format
- ○ I convert from a single source into multiple standard formats
- ○ I convert data from multiple sources into a singe standard format
- ○ I convert data between multiple intermediate formats before I convert into a final standard format
- ○ I am not sure of the conversion process
- ○ Other

If you can, please name the data formats that you regularly use in your research:

In your current research environment, ***data storage*** is (Check all that apply)

- ☐ Offered to you free of charge by your school or institution
- ☐ Offered to you for a fee by your school or institution
- ☐ Created and funded by your department, your lab, or your research group
- ☐ Created and funded through your grants
- ☐ Not sure
- ☐ Other

In your current research environment, ***data management*** is (Check all that apply)

- ☐ Offered to you free of charge by your school or institution
- ☐ Offered to you for a fee by your school or institution
- ☐ Created and funded by your department, your lab, or your research group
- ☐ Created and funded through your grants
- ☐ Not sure
- ☐ Other

<< >>

Data Practices Survey

INDIANA UNIVERSITY

In your current research environment, your ***computing environment*** is (Check all that apply)

- ☐ Offered to you free of charge by your school or institution
- ☐ Offered to you for a fee by your school or institution
- ☐ Created and funded by your department, your lab, or your research group
- ☐ Created and funded through your grants
- ☐ Not sure
- ☐ Other

In your current research environment, is your ***data managed*** by (Check all that apply)

- ☐ A dedicated professional data manager or systems administrator
- ☐ Each individual who creates the data
- ☐ A dedicated graduate assistant or other student
- ☐ A combination of student help and each individual researcher
- ☐ Not sure
- ☐ Other

At what point in your research process does data management become important to you?

- ○ Managing data is not important to my research.
- ○ When the data is created.
- ○ When the analysis begins.
- ○ When the analysis is complete.
- ○ When papers are being written.
- ○ When the data needs to be archived.
- ○ When I need to find something and can't remember where it is.
- ○ Not sure

In the past 5 years, have you lost important data due to (Check all that apply)

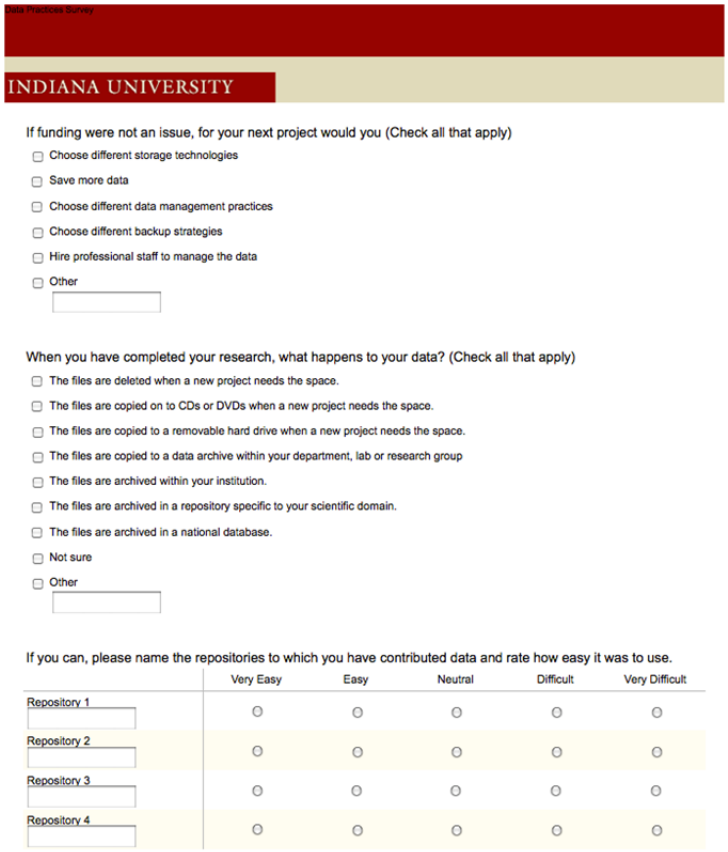
- ☐ Lack of funding
- ☐ Inadvertent human error
- ☐ Malicious hacking
- ☐ Mistakenly thought data was no longer needed
- ☐ Equipment malfunction
- ☐ Lost media
- ☐ Mislabeled media
- ☐ Equipment obsolescence
- ☐ Software no longer recognizes data
- ☐ Physical disaster (flooding, power surges, etc)
- ☐ Data corruption
- ☐ I have not lost data
- ☐ Other

Do you follow standard best practice for backing up your data?

- ○ Yes, almost always
- ○ Sometimes
- ○ Not generally
- ○ No, almost never
- ○ Not sure what is best practice for data backup
- ○ Other

<< >>

Survey Powered By Qualtrics



Data Practices Survey

INDIANA UNIVERSITY

If funding were not an issue, for your next project would you (Check all that apply)

- ☐ Choose different storage technologies
- ☐ Save more data
- ☐ Choose different data management practices
- ☐ Choose different backup strategies
- ☐ Hire professional staff to manage the data
- ☐ Other

When you have completed your research, what happens to your data? (Check all that apply)

- ☐ The files are deleted when a new project needs the space.
- ☐ The files are copied on to CDs or DVDs when a new project needs the space.
- ☐ The files are copied to a removable hard drive when a new project needs the space.
- ☐ The files are copied to a data archive within your department, lab or research group
- ☐ The files are archived within your institution.
- ☐ The files are archived in a repository specific to your scientific domain.
- ☐ The files are archived in a national database.
- ☐ Not sure
- ☐ Other

If you can, please name the repositories to which you have contributed data and rate how easy it was to use.

	Very Easy	Easy	Neutral	Difficult	Very Difficult
Repository 1	○	○	○	○	○
Repository 2	○	○	○	○	○
Repository 3	○	○	○	○	○
Repository 4	○	○	○	○	○

Please indicate the reason that you contributed data to each repository

	Mandated by the journal in which you published	Mandated by your research institution	Mandated by your funding agency	Standard practice in your lab or research group	Individual initiative	Other
Repository 1	○	○	○	○	○	○
Repository 2	○	○	○	○	○	○
Repository 3	○	○	○	○	○	○
Repository 4	○	○	○	○	○	○

If you have deposited research data into a repository, were you able to find it and gain access to it?

- ☐ I have not deposited data into a repository
- ☐ I have not tried to find and access my data in a repository.
- ☐ I was able to find the data and access the data easily.
- ☐ I was able to find the data and access it with some amount of effort.
- ☐ I was able to find the data and access it with a great deal of effort.
- ☐ I was not able to find it.

<< >>

Survey Powered By Qualtrics

INDIANA UNIVERSITY

The following section will ask questions about the metadata you have for your research data. Metadata is the data that is captured to provide information about your research data such as the purpose of the data, the creator of the data, the processes used to create the data, the processes used to normalize, clean and integrate the data, etc.

How often do you have information about your data that is not captured in metadata?

- ○ Almost always
- ○ Sometimes
- ○ Not generally
- ○ Almost never
- ○ Not sure
- ○ Other

How often do you have sufficient metadata to provide all of the information needed to help you and others find your data at a later date?

- ○ Almost always
- ○ Sometimes
- ○ Not generally
- ○ Almost never
- ○ Not sure
- ○ Other

Is your metadata (check all the apply)

- ☐ Stored in a database
- ☐ Stored in a spreadsheet
- ☐ Written in your lab notebook
- ☐ Documented in a text or word processing file
- ☐ Inferred by the file name and directory structure of the data files
- ☐ Not sure
- ☐ Other

Do you use one or more standard metadata formats?

- ○ Yes. If you can, please list formats you use
- ○ No
- ○ Not sure
- ○ Other

<< >>

INDIANA UNIVERSITY

To help your data be more useful to you and to others in the future, how much time would you be willing to spend to create more metadata for your data?

- Up to 10 minutes
- Up to 20 minutes
- More than 20 minutes
- None
- Other

In future projects, would you consider hiring a data professional (e.g. a data librarian or data curator) to help you, your research group, or your lab create better metadata?

- Yes
- Perhaps
- No
- Other

Do you have a formalized set of criteria for judging the quality of your data?

- Yes, almost always
- Sometimes
- Not generally
- No, almost never
- Not sure
- Other

How important is your data quality control process to the quality your research?

- Very important
- Somewhat important
- Not very important
- Not at all important
- Not sure
- Other

Please feel free to provide any additional information about your research, your data, your research process, or the environment in which you do your research.

<< >>

Appendix E. Study Information Sheet

IRB Study #1010002804

Digital Data Research Environments Study

You are invited to participate in a research study to investigate the current data practices of researchers in the United States. Approximately 8,500 researchers have been asked to participate. We ask that you read this form and ask any questions you may have before agreeing to be in the study. If you want to participate, click the URL in the email invitation.

The study is being conducted by Stacy Kowalczyk, a Ph.D. candidate with faculty support of Katy Börner through the School of Library and Information Science at Indiana University.

STUDY PURPOSE

The purpose of this study is to develop a deeper understanding of the role that data plays in the current research environment by developing a data lifecycle model within the research process that accounts for data management, metadata and use of data repositories. This survey will help to refine, extend, and generalize the model.

PROCEDURES FOR THE STUDY:

If you agree to be in the study, you will be asked a series of questions about your current research environment via a web survey. This survey will allow you to skip any questions you do not wish to answer. This survey should take approximately 20 minutes to complete.

No personal or identifying information will be stored. All results will be presented in aggregated formats further obscuring any individual response.

CONFIDENTIALITY

Efforts will be made to keep your personal information confidential. We cannot guarantee absolute confidentiality. Your personal information may be disclosed if required by law. Your identity will be held in confidence in reports in which the study may be published. Organizations that may inspect and/or copy your research records for quality assurance and data analysis include groups such as the study investigator and his/her research associates, the IUB Institutional Review Board or its designees, and (as allowed by law) state or federal agencies, specifically the Office for Human Research Protections (OHRP).

CONTACTS FOR QUESTIONS OR PROBLEMS

If you have any questions about this study or its procedures, please contact Stacy Kowalczyk (skowalcz@indiana.edu). For questions about your rights as a research participant or to discuss

problems, complaints or concerns about a research study, or to obtain information, or offer input, contact the IUB Human Subjects office, 530 E Kirkwood Ave, Carmichael Center, 203, Bloomington IN 47408, 812-856-4242 or by email at iub_hsc@indiana.edu

PAYMENT

You will not receive payment for taking part in this study.

VOLUNTARY NATURE OF STUDY

Taking part in this study is voluntary. You may choose not to take part or may leave the study at any time. Leaving the study will not result in any penalty or loss of benefits to which you are entitled. Your decision whether or not to participate in this study will not affect your current or future relations with the investigator(s).

Form date: November 5, 2010

Appendix F. Data Analysis Plan

Domain	Funding	Size	type of ?	Survey #	Question
category	likert	category		Q2	Which term below best describes your position?
				Q3	What is your primary research institution?
chi sq test	anova	chi sq test	category	Q4	What is your scientific domain? (Please select the domain that most closely describes your research....
anova	correlation	anova	likert	Q5	How is your research funded? Renumbered responses to make a likert-like scale
chi sq test	anova	chi sq test	category	Q6	What is the size of your research group or lab?
chi sq test	anova	chi sq test	multiple answers	Q9_1	What research methods did you use in your project? [check all that apply]-Surveys
chi sq test	anova	chi sq test		Q9_2	What research methods did you use in your project? [check all that apply]-Field studies
chi sq test	anova	chi sq test		Q9_3	What research methods did you use in your project? [check all that apply]-Case studies
chi sq test	anova	chi sq test		Q9_4	What research methods did you use in your project? [check all that apply]-Direct observation in experimental situations
chi sq test	anova	chi sq test		Q9_5	What research methods did you use in your project? [check all that apply]-Analysis of instrument generated data
chi sq test	anova	chi sq test		Q9_6	What research methods did you use in your project? [check all that apply]-Analysis of existing data sets
chi sq test	anova	chi sq test		Q9_7	What research methods did you use in your project? [check all that apply]-Modeling and simulation
chi sq test	anova	chi sq test		Q9_8	What research methods did you use in your project? [check all that apply]-Text or language analysis
anova	correlation	anova	likert	Q10	Do you worry about the longevity of your data?

anova	correlation	anova	likert	Q11	The best term to describe your level of concern about preserving your data is
chi sq	anova	chi sq	Y/N	Q12	Do you have any contractual obligations (through grants or other agreements) to keep your data usabl...
anova	correlation	anova	likert	Q13	Is it important to you to make your research data available to future generations of researchers?
anova	correlation	anova	likert	Q14	For your research data, how easy is it for you to identify the most important data to preserve?
anova	correlation	anova	likert	Q15	For your research data, how easy is it for you to identify the data that is most in need of preserva...
chi sq test	anova	chi sq test	multiple answers	Q16_1	In your research, do you use: (Check all that apply)-Data that you have **created** from observation, instruments, experiments, or other processes
chi sq test	anova	chi sq test		Q16_2	In your research, do you use: (Check all that apply)-Data that you **gathered** from other sources such as databases, vendors, or webcrawls
chi sq test	anova	chi sq test	multiple answers	Q17_1	Which of the following processes do you run on your data? (Check all that apply)-Data normalizing (resolving scale issues, reformatting for consistency, etc.)
chi sq test	anova	chi sq test		Q17_2	Which of the following processes do you run on your data? (Check all that apply)-Data cleaning (fixing errors)
chi sq test	anova	chi sq test		Q17_3	Which of the following processes do you run on your data? (Check all that apply)-Data integration (merging data from several sources)
chi sq test	anova	chi sq test		Q17_4	Which of the following processes do you run on your data? (Check all that apply)-Instrument calibration

chi sq test	anova	chi sq test	category	Q18	For the project you identified earlier, how much time is spent on the data normalization, cleaning,...
chi sq test	anova	chi sq test	multiple answers	Q19_1	After the data is collected and any data normalization, cleaning and integration processes for the...-I have observation data that is unique
chi sq test	anova	chi sq test		Q19_2	After the data is collected and any data normalization, cleaning and integration processes for the...-I have experimental data that is unique
chi sq test	anova	chi sq test		Q19_3	After the data is collected and any data normalization, cleaning and integration processes for the...-Data is unique due to the quantity and quality of the data
chi sq test	anova	chi sq test		Q19_4	After the data is collected and any data normalization, cleaning and integration processes for the...-Data is unique due to the level of uniformity and integration of the data
chi sq test	anova	chi sq test		Q19_5	After the data is collected and any data normalization, cleaning and integration processes for the...-Data is unique due to the longitudinal nature of the data
chi sq test	anova	chi sq test		Q19_6	After the data is collected and any data normalization, cleaning and integration processes for the...-Data is unique due to the added value of metadata
chi sq test	anova	chi sq test		Q19_7	After the data is collected and any data normalization, cleaning and integration processes for the...-Data is not unique and can be recreated from the original sources
chi sq test	anova	chi sq test		Q19_10	After the data is collected and any data normalization, cleaning and integration processes for the...-Data is unique due to the integration of unique analysis into the data
chi sq test	anova	chi sq test	category	Q20	In the project you identified earlier, approximately how many times did you convert data from one fo...
chi sq test	anova	chi sq test	category	Q21	For the project you identified earlier, which of the scenarios below would best

					describe the format...
chi sq test	anova	chi sq test	multiple answers	Q23_1	In your current research environment, data storage is (Check all that apply) -Offered to you free of charge by your school or institution
chi sq test	anova	chi sq test		Q23_2	In your current research environment, data storage is (Check all that apply) -Offered to you for a fee by your school or institution
chi sq test	anova	chi sq test		Q23_3	In your current research environment, data storage is (Check all that apply) -Created and funded by your department, your lab, or your research group
chi sq test	anova	chi sq test		Q23_4	In your current research environment, data storage is (Check all that apply) -Created and funded through your grants
chi sq test	anova	chi sq test	multiple answers	Q24_1	In your current research environment, data management is (Check all that apply) -Offered to you free of charge by your school or institution
chi sq test	anova	chi sq test		Q24_2	In your current research environment, data management is (Check all that apply) -Offered to you for a fee by your school or institution
chi sq test	anova	chi sq test		Q24_3	In your current research environment, data management is (Check all that apply) -Created and funded by your department, your lab, or your research group
chi sq test	anova	chi sq test		Q24_4	In your current research environment, data management is (Check all that apply) -Created and funded through your grants
chi sq test	anova	chi sq test	multiple answers	Q25_1	In your current research environment, your computing environment is (Check all that apply) -Offered to you free of charge by your school or institution
chi sq test	anova	chi sq test		Q25_2	In your current research environment, your computing environment is (Check all that apply) -Offered to you for a fee by your school or institution

chi sq test	anova	chi sq test		Q25_3	In your current research environment, your computing environment is (Check all that apply) -Created and funded by your department, your lab, or your research group
chi sq test	anova	chi sq test		Q25_4	In your current research environment, your computing environment is (Check all that apply) -Created and funded through your grants
chi sq test	anova	chi sq test	multiple answers	Q26_1	In your current research environment, is your data managed by (Check all that apply)-A dedicated professional data manager or systems administrator
chi sq test	anova	chi sq test		Q26_2	In your current research environment, is your data managed by (Check all that apply)-Each individual who creates the data
chi sq test	anova	chi sq test		Q26_3	In your current research environment, is your data managed by (Check all that apply)-A dedicated graduate assistant or other student
chi sq test	anova	chi sq test		Q26_4	In your current research environment, is your data managed by (Check all that apply)-A combination of student help and each individual researcher
chi sq test	anova	chi sq test	category	Q27	At what point in your research process does data management become important to you?
chi sq test	anova	chi sq test	multiple answers	Q28_1	In the past 5 years, have you lost important data due to (Check all that apply) -Lack of funding
chi sq test	anova	chi sq test		Q28_2	In the past 5 years, have you lost important data due to (Check all that apply) -Inadvertent human error
chi sq test	anova	chi sq test		Q28_3	In the past 5 years, have you lost important data due to (Check all that apply) -Malicious hacking
chi sq test	anova	chi sq test		Q28_4	In the past 5 years, have you lost important data due to (Check all that apply) -Mistakenly thought data was no longer needed

chi sq test	anova	chi sq test		Q28_5	In the past 5 years, have you lost important data due to (Check all that apply) -Equipment malfunction
chi sq test	anova	chi sq test		Q28_6	In the past 5 years, have you lost important data due to (Check all that apply) -Lost media
chi sq test	anova	chi sq test		Q28_7	In the past 5 years, have you lost important data due to (Check all that apply) -Mislabeled media
chi sq test	anova	chi sq test		Q28_8	In the past 5 years, have you lost important data due to (Check all that apply) -Equipment obsolescence
chi sq test	anova	chi sq test		Q28_9	In the past 5 years, have you lost important data due to (Check all that apply) -Software no longer recognizes data
chi sq test	anova	chi sq test		Q28_10	In the past 5 years, have you lost important data due to (Check all that apply) -Physical disaster (flooding, power surges, etc)
chi sq test	anova	chi sq test		Q28_11	In the past 5 years, have you lost important data due to (Check all that apply) -Data corruption
chi sq test	anova	chi sq test		Q28_12	In the past 5 years, have you lost important data due to (Check all that apply) -I have not lost data
chi sq test	anova	chi sq test	Y/N/U/O	Q29	Do you follow standard best practice for backing up your data?
chi sq test	anova	chi sq test	multiple answers	Q30_1	If funding were not an issue, for your next project would you (Check all that apply)-Choose different storage technologies
chi sq test	anova	chi sq test		Q30_2	If funding were not an issue, for your next project would you (Check all that apply)-Save more data
chi sq test	anova	chi sq test		Q30_3	If funding were not an issue, for your next project would you (Check all that apply)-Choose different data management practices

chi sq test	anova	chi sq test		Q30_4	If funding were not an issue, for your next project would you (Check all that apply)-Choose different backup strategies
chi sq test	anova	chi sq test		Q30_5	If funding were not an issue, for your next project would you (Check all that apply)-Hire professional staff to manage the data
chi sq test	anova	chi sq test	multiple answers	Q31_1	When you have completed your research, what happens to your data? (Check all that apply)-The files are deleted when a new project needs the space.
chi sq test	anova	chi sq test		Q31_2	When you have completed your research, what happens to your data? (Check all that apply)-The files are copied on to CDs or DVDs when a new project needs the space.
chi sq test	anova	chi sq test		Q31_3	When you have completed your research, what happens to your data? (Check all that apply)-The files are copied to a removable hard drive when a new project needs the space.
chi sq test	anova	chi sq test		Q31_4	When you have completed your research, what happens to your data? (Check all that apply)-The files are copied to a data archive within your department, lab or research group
chi sq test	anova	chi sq test		Q31_5	When you have completed your research, what happens to your data? (Check all that apply)-The files are archived within your institution.
chi sq test	anova	chi sq test		Q31_6	When you have completed your research, what happens to your data? (Check all that apply)-The files are archived in a repository specific to your scientific domain.
chi sq test	anova	chi sq test		Q31_7	When you have completed your research, what happens to your data? (Check all that apply)-The files are archived in a national database.
anova	correlation	anova	likert	Q32	If you can, please name the repositories to which you have contributed data and rate how easy it was...-Repository 1

chi sq test	anova	chi sq test	category	Q33	Please indicate the reason that you contributed data to each repository-Repository 1
chi sq test	anova	chi sq test	category	Q34	If you have deposited research data into a repository, were you able to find it and gain access to...-I have not deposited data into a repository
anova	correlation	anova	likert	Q36	How often do you have information about your data that is not captured in metadata?
anova	correlation	anova	likert	Q37	How often do you have sufficient metadata to provide all of the information needed to help you and o...
chi sq test	anova	chi sq test	multiple answers	Q38_1	Is your metadata (check all the apply)-Stored in a database
chi sq test	anova	chi sq test		Q38_2	Is your metadata (check all the apply)-Stored in a spreadsheet
chi sq test	anova	chi sq test		Q38_3	Is your metadata (check all the apply)-Written in your lab notebook
chi sq test	anova	chi sq test		Q38_4	Is your metadata (check all the apply)-Documented in a text or word processing file
chi sq test	anova	chi sq test		Q38_5	Is your metadata (check all the apply)-Inferred by the file name and directory structure of the data files
chi sq test	anova	chi sq test	Y/N/U/O	Q39	Do you use one or more standard metadata formats?
chi sq test	anova	chi sq test	categories	Q40	To help your data be more useful to you and to others in the future, how much time would you be will...
chi sq test	anova	chi sq test	Y/P/N/O	Q41	In future projects, would you consider hiring a data professional (e.g. a data librarian or data cur...
anova	correlation	anova	likert	Q42	Do you have a formalized set of criteria for judging the quality of your data?

anova	correlation	anova	likert	Q43	How important is your data quality control process to the quality your research?

Appendix G. Uniqueness by Demographic Categories

Uniqueness and Domain

	More Likely	**Less Likely**
I have observation data that is unique (χ^2_7 = 64.905, p < .001)	Biology Education Geoscience	Computer Science Engineering Mathematics Physical Science
I have experimental data that is unique (χ^2_7 = 97.200, p < .001)	Biology Physical Science	Education Geoscience Mathematics Social Science
Data is unique due to the quantity and quality of the data (χ^2_7 = 45.940, p < .001)	Biology Geoscience	Computer Science Mathematics
Data is unique due to the level of uniformity and integration of the data (χ^2_7 = 28.264, p < .001)	Biology Geoscience Social Science	Education Engineering Mathematics
Data is unique due to the longitudinal nature of the data (χ^2_7 = 29.488, p < .001)	Biology Education Social Science	Mathematics Physical Science
Data is unique due to the added value of metadata (χ^2_7 = 44.238, p < .001)	Biology	Engineering Mathematics Physical Science
Data is not unique and can be recreated from the original sources (χ^2_7 = 6.759, p = .454)	n/a	n/a
Data is unique due to the integration of unique analysis into the data (χ^2_7 = 17.098, p < .017)	Biology Computer Science Geoscience	Mathematics

Uniqueness and Size of Lab

	More Likely	Less Likely
I have observation data that is unique (χ_2^2 = 2.878, p = .237)	n/a	n/a
I have experimental data that is unique (χ_2^2 = 82.938, p < .001)	Large Labs	Individual Mid-Sized Labs
Data is unique due to the quantity and quality of the data (χ_2^2 = 13.013, p = .001)	Large Labs	Individual Mid-Sized Labs
Data is unique due to the level of uniformity and integration of the data (χ_2^2 = 12.011, p = .002)	Large Labs	Individual Mid-Sized Labs
Data is unique due to the longitudinal nature of the data (χ_2^2 = .238, p =. 888)	n/a	n/a
Data is unique due to the added value of metadata (χ_2^2 = 20.592, p < .001)	Large Labs	Individual Mid-Sized Labs
Data is not unique and can be recreated from the original sources (χ_2^2 = 1.965, p = .374)	n/a	n/a
Data is unique due to the integration of unique analysis into the data (χ_2^2 = 6.418, p = .040)	Large Labs	Individual Mid-Sized Labs

Uniqueness and Funding Source

I have observation data that is unique ($F_{4,785} = 2.914, p = .018$)	The more grant funding the higher the use of observational data
I have experimental data that is unique ($F_{4,785} = 2.573, p = .035$)	Both ends of the funding spectrum are more likely to have experimental data
Data is unique due to the quantity and quality of the data ($F_{4,785} = 2.778, p = .020$)	Both ends of the funding spectrum are more likely to do this
Data is unique due to the level of uniformity and integration of the data ($F_{4,785} = .308, p = .690$)	n/a
Data is unique due to the longitudinal nature of the data ($F_{4,785} = .576, p = .402$)	n/a
Data is unique due to the added value of metadata ($F_{4,785} = 1.246, p = .044$)	Exclusively or mostly grant funded more likely to report this.
Data is not unique and can be recreated from the original sources ($F_{4,785} = .615, p = .297$)	n/a
Data is unique due to the integration of unique analysis into the data ($F_{4,785} = .485, p = .425$)	n/a

Appendix H. Technical Environment Constructs by Demographic Categories

Data Storage by Size of Lab

	More Likely	Less Likely
Offered free by your school or institution ($\chi^2_2 = 13.328$, p = .001)	Individuals Mid-size Labs	Large Labs
Offered for a fee by your school or institution ($\chi^2_2 = 10.321$, p = .006)	Large Labs	Individuals Mid-size Labs
Created and funded by your department ($\chi^2_2 = 43.837$, p < .001)	Large Labs	Individuals Mid-size Labs
Created and funded through your grants ($\chi^2_2 = 35.037$, p < .001)	Large Labs	Individuals Mid-size Labs

Data Storage by Scientific Domain

	More Likely	Less Likely
Offered free by your school or institution ($\chi^2_7 = 21.464$, p = .003)	Social Science	Computer Science Geosciences Physical Science
Offered for a fee by your school or institution ($\chi^2_7 = 9.383$, p = .226)	n/a	n/a
Created and funded by your department ($\chi^2_7 = 23.000$, p = .002)	Geosciences Physical Science	Education Mathematics
Created and funded through your grants ($\chi^2_7 = 61.245$, p < .001)	Biology Geosciences Physical Science	Computer Science Engineering Mathematics Social Science

Data Storage and Funding

Offered free by your school or institution ($F_{4,785} = 4.671, p < .001$)	The more institutional funding the more people had access to free storage.
Offered for a fee by your school/institution ($F_{4,785} = .165, p = .730$)	n/a
Created and funded by your department ($F_{4,785} = 2.363, p = .016$)	Researchers with mixed funding were more likely to report this.
Created and funded through your grants ($F_{4,785} = 9.691, p < .001$)	Exclusively and primarily grant funded are more likely to report this.

Data Management by Size of Lab

	More Likely	**Less Likely**
Offered free by your school or institution ($\chi^2_2 = .363$, p = .834)	n/a	n/a
Offered for a fee by your school or institution ($\chi^2_2 = 2.527$, p = .283)	n/a	n/a
Created and funded by your department ($\chi^2_2 = 30.705$ p < .001)	Large Labs	Individuals Mid-sized Labs
Created and funded through your grants ($\chi^2_2 = 12.084$, p = .002)	Large Labs	Individual

Data Management by Scientific Domain

	More Likely	Less Likely
Offered free by your school or institution (χ^2_7 = 7.943, p = .338)	n/a	n/a
Offered for a fee by your institution (χ^2_7 = 10.297, p = .172)	n/a	n/a
Created and funded by your department (χ^2_7 = 20.559, p = .004)	Computer Science	Engineering Mathematics
Created and funded through your grants (χ^2_7 = 51.306, p < .001)	Biology Physical Science	Computer Science Engineering Mathematics

Data Management and Funding

Offered free by your school or institution (*$F_{4,785}$ = .612, p = .273*)	n/a
Offered for a fee by your school /institution (*$F_{4,785}$ = .145, p = .477*)	n/a
Created and funded by your department (*$F_{4,785}$ = 4.099, p < .001*)	Researchers with mixed funding were more likely to report this.
Created and funded through your grants (*$F_{4,785}$ = 12.148, p < .001*)	Exclusively and primarily grant funded are more likely to report this.

Computing Environment by Size of Lab

	More Likely	**Less Likely**
Offered free by your school or institution (χ_2^2 = 27.193, p = .001)	Individual Mid-sized Labs	Large Labs
Offered for a fee by your school or institution (χ_2^2 = 7.418, p = .025)	Large Labs	Mid-sized Labs
Created and funded by your department (χ_2^2 = 38.148, p > .001)	Large Labs	Individual Mid-sized Labs
Created and funded through your grants (χ_2^2 = 28132, p > .001)	Large Labs	Individual

Computing Environment by Scientific Domain

	More Likely	**Less Likely**
Offered free by your school or institution (χ_7^2 = 25.273, p = .001)	Mathematics Social Science	Computer Science Engineering Physical Science
Offered for a fee by your school/institution (χ_7^2 = 11.665, p = .112)	n/a	n/a
Created and funded by your department (χ_7^2 = 25.895, p = .001)	Geosciences Physical Sciences	Computer Science Education Mathematics Social Sciences
Created and funded through your grants (χ_7^2 = 58.195, p < .001)	Biology Geoscience Physical Science	Computer Science Engineering Mathematics Social Science

Computing Environment and Funding

Offered free by your school or institution ($F_{4,785} = 3.903, p = .003$)	Researchers with mixed funding were more likely to report this.
Offered for a fee by your school /institution ($F_{4,785} = .300, p = .569$)	n/a
Created and funded by your department, ($F_{4,785} = 1.391, p = .178$)	n/a
Created and funded through your grants ($F_{4,785} = 7.664, p < .001$)	Exclusively and primarily grant funded are more likely to report this.

Vita

Stacy Kowalczyk has spent most of her professional life working in library automation. She managed software development at NOTIS Systems, a library information system vendor; and she managed the development of the technical infrastructure for Harvard's Library Digital Initiative. During her Ph.D. work, she held the position of Associate Director for Projects and Services for the Indiana University Libraries Digital Library Program. She has an undergraduate degree in English Literature from Lewis University, a Masters of Library and Information Science from Dominican University, and a Ph.D. from Indiana University's School of Library and Information Science.

Recent Publications

Kowalczyk, S.T. (2011). Towards a model of the e-Science Data Environment. In the Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries, Ottawa Canada, June 13-17, 2011.

Kowalczyk, S.T. & Shankar, K. (2011). Data sharing in the sciences. In B. Cronin (Ed.), Annual Review of Information Science and Technology, Volume 45 (2011), pp. 247 - 294.

Dennis, A.R., Robert, L.P., Curtis, A.M., Kowalczyk, S.T. & Hasty, B.K. (2011). Trust is in the eye of the beholder: A vignette study of post-event behavioral controls' effects on individual trust in virtual teams. *Information Systems Research* 22 (2), 213 – 417.

Kowalczyk, S.T. (2007). Digital preservation by design. In M. Raisinghani (Ed.) *Handbook of Research on Global Information Technology Management in the Digital Economy*. New York: IGI Publishing. Pre-publication version available at http://ella.slis.indiana.edu/~skowalcz/Book_Chapter/Dig_Pres_by_Design.pdf

Brancolini, K., Kowalczyk, S.T. &Riley, J. (2006). IN Harmony: Sheet music from Indiana. *First Monday, 11(8).* Available from http://www.firstmonday.org/issues/issue11_8/brancolini/

Recent Presentations and Invited Talks

Kowalczyk, S.T. (2010). Data publishing. Presented at *American Society for Information Science and Technology Research Data Access and Preservation Summit,* April 9-10, 2010, Phoenix, AZ. Presenter and Discussant. Invited Talk.

Kowalczyk, S.T. (2009). Libraries in Bamboospace. *Bamboo Workshop 3,* January 12 - 14, 2009, Tucson, Arizona. Invited Talk.

Kowalczyk, S.T. & Halliday. J. (2008). A Multi-tiered architecture for distributed data collection and centralized data delivery. *The Digital Library Federation Spring Forum,* April 28, 2008, Minneapolis, Minnesota.