IJDC | *Peer-Reviewed Paper*

# Modelling the Digital Research Data Lifecycle

Stacy T. Kowalczyk
Dominican University

## Abstract

This paper develops and tests a lifecycle model for the preservation of research data by investigating the research practices of scientists. This research is based on a mixed-method approach. An initial study was conducted using case study analytical techniques; insights from these case studies were combined with grounded theory in order to develop a novel model of the Digital Research Data Lifecycle. A broad-based quantitative survey was then constructed to test and extend the components of the model. The major contribution of these research initiatives is the creation of the Digital Research Data Lifecycle, a data lifecycle that provides a generalized model of the research process to better describe and explain both the antecedents and barriers to preservation. The antecedents and barriers to preservation are data management, contextual metadata, file formats, and preservation technologies. The availability of data management support and preservation technologies, the ability to create and manage contextual metadata, and the choices of file formats all significantly affect the quality of research data preserved.

*Received* 28 August 2016 ~ *Revision Received 5* December 2017 ~ *Accepted* 8 May 2018

Correspondence should be addressed to Stacy T. Kowalczyk, 7900 West Division Street, River Forest, IL 60305
Email: skowalczyk@dom.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v12i2.429
DOI: 10.2218/ijdc.v12i2.429

# Introduction

Computer-based technology has greatly increased the quantity of research data. This data deluge is becoming an important area of study in computer science, information science, and domain sciences. One of the primary areas of study in this emerging field is data preservation; that is, the ability to provide long term access to the data for reuse. 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 (Association of Research Libraries [ARL], 2006; National Science Board [NSB], 2005; Tibbo, 2014); it is an integral part of the scientific record as evidence of the rhetorical structure of scholarly communication (Rusbridge, 2007; Swan and Brown, 2008); it is necessary for replication and validation of scientific results (Swan and Brown, 2008); the data has significant economic value as intellectual capital, an important and invaluable resource that can be used repeatedly (ARL, 2006; Rumsey, 2006); and scientific digital data is a generalized good in that society benefits both directly and indirectly when this data is available for citizen scientists, for teaching, for commercial reuse, and for policy development (Lord and Macdonald, 2003; Rumsey, 2010).

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 and Wheeler, 2007). However, there is a lack of "evidence from the community of active researchers with respect to their own needs and aspirations within the research life cycle" about data management roles and responsibilities (Pryor and Donnelly, 2009). New research is needed to describe, measure, and mitigate the "obstacles to the longevity of digital materials" (Ross, 2007).

There is surprisingly little quantitative data describing the research 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, Weaver, Bradbury, and Porter, 2008; Jones, Ball, and Ekmekcioglu, 2008; Lyon, 2007; Marcus, Ball, Delserone, Hribar, and Loftus, 2007; Pinfield, Cox and Smith, 2014; Pritchard, Anand, and Carver, 2005; Pryor, 2007; Witt, Carlson, Brandt, and Cragin, 2009). Little of this data was used to develop theory. Theory development backed by quantitative data is necessary for the field to progress.

This research uses both qualitative and quantitative data to develop a new model for describing the lifecycle of research data. This research provides new insights into the workflow of the digital science process by quantifying the current state of digital science data management and describing the multiple environments in which data is created, used, saved, and preserved. The research questions that frame this study are:

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. What are the antecedents to preservation?
4. How do the threats to preservation affect the data lifecycle?

The first section of the paper will set the theoretical foundations for this research by reviewing the current and prior work on digital preservation of scientific data from the computer science, information science, and domain sciences literatures. The methodology for this research follows the theoretical foundations. The results section describes the outcomes of the studies. The discussion of the results will develop the theoretical model of the Digital Research Data Lifecycle. The concluding section describes the impacts and implications of the presented research.

# Theoretical Foundations

### 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, 2007; Tibbo, 2014). Data lifecycles are path dependent (Rumsey, 2010), meaning that the cumulative weight of decisions made at each stage determines what is available at the next step (Wallis, Borgman, Mayernik and Pepe, 2008). 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).

A number of preservation and curation lifecycles have been developed including the Digital Curation Center's Curation Lifecycle Model (Higgins, 2008), the Library of Congress' Digital Preservation Lifecycles, (Library of Congress, 2011), and the SHAMAN Project's Information Lifecycle (Wilkes, Brunsmann, Heutelbeck, Hundsdörfer, Hemmje, and Heidbrink, 2011). These are generalized views that depict the process of curating existing data but do not have a detailed description of the process of the researchers who produce the data (Higgins, 2008; Library of Congress, 2011). A more general life cycle model has been developed by the UK Data Archive. This model describes the steps to create research data including planning research, collecting data, processing and analyzing data, publishing and sharing data, preserving data, and reusing data (UK Data Service, 2018). This is depicted as a unidirectional cycle starting with data creation. The model addresses data sharing, data formats and migration, data licensing and other issues that are of vital importance to data archival repositories.

Within the literature, many data lifecycles that describe the process of creating data are either generic in that they pertain to an entire domain, or they are overly specific because they pertain to one particular lab or project. Domain or project-specific lifecycles have been developed from case studies (Borgman, 2007; Higgins, 2008; Long, Mantey, Wittenbrink, Haining, and Montague, 1995; Wallis et al., 2008). Lifecycles differ widely between different scientific domains (Borgman, Bowker, Finholt, and Wallis, 2009). Data management, data preservation, or data archiving are usually depicted as the final steps in the lifecycle of data. Data preparation for preservation or archival activities, intermediate file management, and assessment criteria for determining which files to preserve, are not discussed in existing models.

Institutions, data centers, users, funders, data creators, and publishers all have roles, rights, and responsibilities for curating, archiving, and preserving data (Hey and 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 for managing data for the life of the project, meeting standards of good practice, and for "work[ing] up data" for use by others (Lyon, 2007). The burden is on the scientist to manage the data, create the contextual metadata, and determine a final disposition of the data.

## 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 and Davidson, 2009; Rice, 2007; Zilinski, Scherer, Bullock, Horton, and Matthews, 2014). As the amount of data increases, so does the complexity and resources required to manage the data (Pritchard, Anand, and Carver, 2005).

Established and well-funded data collections often have dedicated data management staff (Karasti, Baker, and Halkola, 2006) and large data centers to manage the petabyte datasets (Gray, Liu, Nieto-Santisteban, Szalay, DeWitt, and Heber, 2005; Hey and Trefethen, 2003). However, managing data is most often the responsibility of individual scientists (Clements and McCutcheon, 2014; Henty, Weaver, Bradbury, and Porter, 2008; Lynch, 2008; Lyon, 2008; Pritchard, Anand, and 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, and Carver, 2005; Pryor and Donnelly, 2009; Ray, 2014). Without clear direction from funding agencies, researchers are left to create their own guidelines (Carlson, Johnston, and Huang, 2014; Jones, Ball, and Ekmekcioglu, 2008; Marcus, Ball, Delserone, Hribar, and Loftus, 2007). As discussed above, 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.

## Contextual Metadata

Metadata is a key factor for data preservation (Anderson, 2004; Hey and Trefethen, 2003; NISO Framework Working Group [NISO], 2007; Rajasekar and Moore, 2001; Swan and Brown, 2008), for replicating results in the peer review process (Vardigan, Heus, and Thomas, 2008), for data reuse (Gray et al, 2005; Hey and Trefethen, 2003; Lyon, 2007), and for creating knowledge from data (Hey and Trefethen, 2003). Metadata can be as valuable as the original data (Ray, 2014). 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 (Minor, Critchlow, Hutt, Fleming, Bergstrom, and Sutton, 2014). And metadata is expensive; there is usually a direct relationship between the cost of metadata creation and the benefit to the user (NISO, 2007). Creating metadata is a demanding task that is both complex and time-consuming (Michener, 2006; Minor et al., 2014; Pryor, 2007). Determining which attributes to document, that is what to capture in metadata, is a difficult task for researchers (Ray, 2014). 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).

### Formats

Scientific data, be it numeric, text-based, audio, video, or still images, must be represented in a format. A format is defined as “the internal structure and encoding of a digital object, which allows it to be processed, or to be rendered in a human-accessible form” (Brown, 2006). As the representation of data, format becomes an antecedent to preservation by providing the knowledge of the data itself. During the research life cycle, the original, raw data is often processed and transformed into a reduced or derived data product (Swan and Brown, 2008). Reading, rendering, and processing the data requires knowledge and understanding of the format. The probability of long-term persistence and access depends on the complexity and transparency of the file format itself; changes to the format or to the environment can make the underlying data unrecognizable and unusable (Abrams, 2004). Many domain-centric communities, such as astrophysics and genomics, have developed sets of standard formats to facilitate data sharing and system interoperability (Gardner et al., 2003; Westbrook, Ito, Nakamura, Henrick, and Berman, 2005); however, the standards are not well understood by the researchers for which they were designed (Mayernik, 2015). Effective data reuse, and ultimately the ability to preserve data, depends on standard data formats and tool sets to build, read and manage the data (Gardner et al., 2003, Ray, 2014).

### Preservation Technologies

Creating a persistent data collection requires a set of technologies that Moore (2008) 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, and 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, persistence, and access (Ray, 2014). 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; Tibbo, 2014). 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.

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). Different research domains have very different requirements and constraints for data infrastructure (Tenopir, Dalton, Allard, Frame, Pjesivac, Birch, et al., 2015). 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).

## Antecedents as Barriers to Preservation

The antecedents to preservation – data management, contextual metadata, formats, 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; Vines, Albert, Andrew, Débarre, Bock, Franklin et al., 2014).

### Data management as a barrier to preservation

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 (Kowalczyk, 2008). With all academic incentives rewarding new work, it is counterintuitive for scientists to invest time, effort, and money to care for older data (Minor, Critchlow, Hutt, Fleming, Bergstrom, and Sutton, 2014). Data management is frequently considered to be overhead and not research (Anderson, 2004) and can be a burden for scientists (Bell, Hey, and Szalay, 2009; Kowalczyk, 2014).

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, and Harboe-Ree, 2007; Kowalczyk, 2014). 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). Without ongoing data management, 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 and Macdonald, 2003).

### Metadata as a 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, and Thomas, 2008). However, the process of creating metadata, what is described as the “mechanics of metadata,” causes both confusion and frustration among researchers (Swan and Brown, 2008). 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.” The incentives for creating usable, machine-processable metadata are not strong enough to overcome this absence of useable tools.

### Formats as barriers

The wide variety of data formats used in different scientific disciplines and the amount of effort to convert data into different formats can create a barrier to preservation (Mann, Williams, Atkinson, Brodlie, Storkey, and Williams, 2002; Shiffrin and Börner, 2004). 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). Lawrence and colleagues (2000) discovered that it was easier to manage data in open formats, such as TIFF. But, they also discovered that even open formats can have a proprietary and thus, secret, set of tags that may not be supported in future versions. Because they manage and archive large collections of heterogeneous digital files, the Library of Congress, the National Archives of England, Wales and the United Kingdom, OCLC, and the National Archives of Australia have developed criteria for assessing the risks associated with formats. While these four sets of criteria differ in levels of complexity and thoroughness, there is a consensus that archival formats should be well documented and well understood; not wholly owned by a single commercial entity; widely adopted to increase the probability of commercial tools for migration and to ensure a long usage cycle to avoid repeated, short term migration; self-contained; and not reliant on a specific technical environment (Kowalczyk, 2008).

### Preservation technology as a 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.

A major barrier for preservation is the complicated, inflexible, and counterintuitive processes required to deposit data within these repositories (van Westrienen and 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.

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, not discoverable, or not recoverable (Kowalczyk, 2008).

# Methodology

This research used both qualitative and quantitative research methodologies. Grounded theory, a qualitative research methodology for inductively deriving theory based on the data gathered about one or more phenomena (Strauss and Corbin, 1990), was used to develop a set of constructs and relationships that describe the lifecycle of digital research data. A quantitative survey was then developed to verify, validate and extend the constructs and relationships developed by the grounded theory process. 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).

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 and 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 and Huberman, 1994; Glaser and Strauss, 1967; Voss, Tsikriktsis, and 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 the grounded theory development portion of this research 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. The data for this study was collected in one-hour semi-structured interviews with the directors of the 11 laboratories and research centers. After iterative coding, a set of six primary constructs emerged: data creation, quality control, content, data collections, context and format. Due to the limited number of researchers interviewed and the nature of the data collected, the interactions between the six constructs were implied and could not be explicitly defined.

The preliminary study sample used the theoretically important categories of lab size, lab funding, and scientific domain. A recent JISC report indicated 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, and 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. 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 and 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.

To gather quantitative data to test, expand, and extend the constructs, a survey was conducted using a broad survey frame of grant awardees of the National Science Foundation (NSF). The potential participant pool of NSF grant awardees from 2007 through 2010 was retrieved from the Scholarly Database (LaRowe, Ambre, Burgoon, Ke, and Börner, 2009). From the set of 41,917 principal investigators in the database, a subset of 1,200 unique PIs from each of the seven NSF directorates was selected randomly for a total pool of 8,400 potential participants. The 8,400 principal investigators were asked to participate via an email solicitation; 897 researchers took the survey, resulting in a response rate of 10.6% for the original sample.

The survey was designed to gather four demographic factors: researcher role, scientific domain, funding source, and size of laboratory. Of the participants in the study, 90% self-identified their role as principal investigators, 4% as Researcher, 2% as Post Doc, and 3% as other (including professor, librarians, or data stewards). As is evident from the numbers, the researcher role did not prove to be a useful demographic for statistical purposes.

The other demographics of the respondents are more evenly distributed than the research roles. The participants represent a diverse set of scientific domains; 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. The domains are distributed though the sample as follows: Biological Sciences (22%), Computer and Information Science (11%), Education (4%), Engineering (13%), Geosciences (13%), Mathematical Sciences (8%), Physical Sciences (17%), and Social, Behavioral and Economic Sciences (13%). A large majority of the respondents, 82%, work 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. The results of this survey indicate that 85% researchers are funded exclusively or primarily from grants. Only 8% of responders are funded exclusively or primarily from their institution. The participants in this study identified affiliations with 334 unique research institutions. 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. 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.

# Results

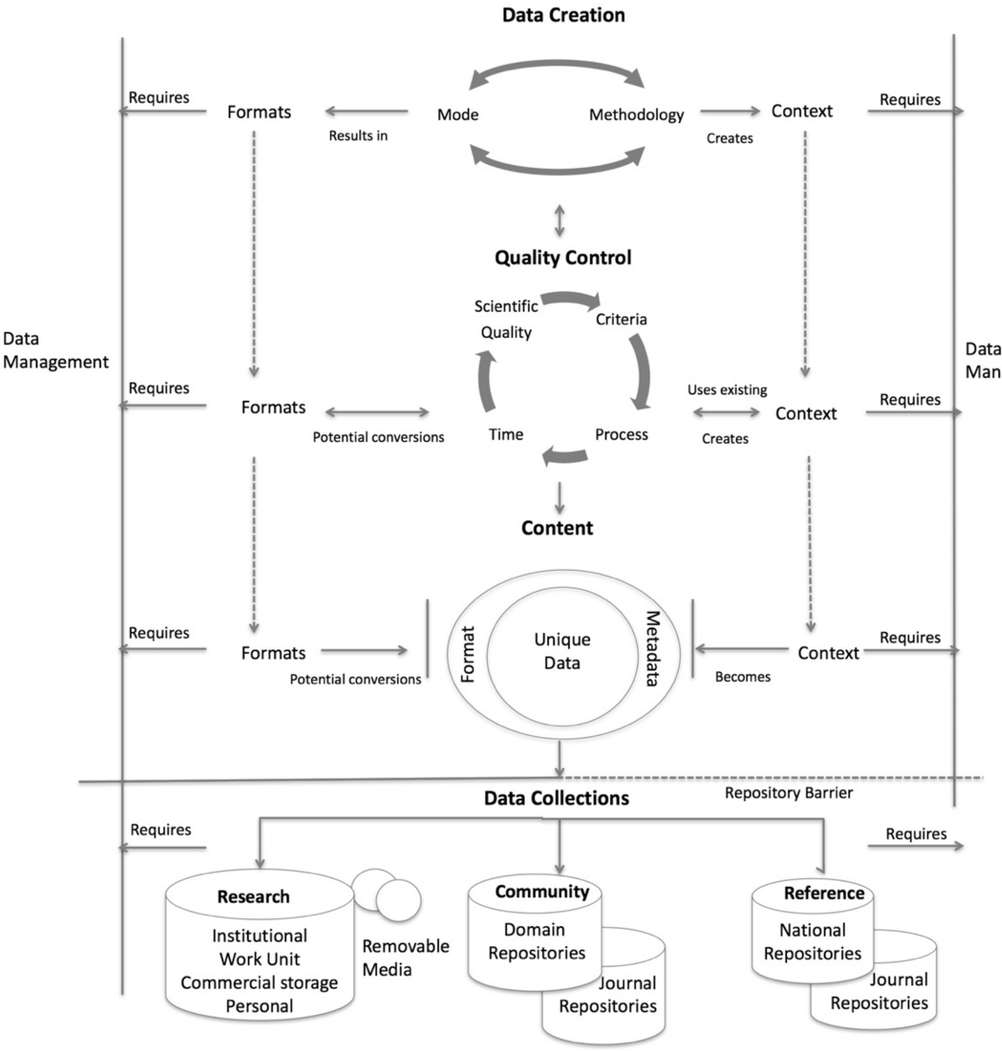

**Figure 1.** The Digital Research Data Lifecycle.

In this section, a theoretical model of the Digital Research Data Lifecycle will be developed. This new lifecycle can be considered theoretical as it “consists of plausible relationships proposed among concepts and sets of concepts” (Strauss and Corbin, 1990). The Digital Research Data Lifecycle (see Figure 1) is based on the data from both the preliminary study interviews and the responses to the survey described above. This detailed model has been constructed with the rich set of qualitative and quantitative data. The model describes a generalized research data lifecycle that accounts for the antecedents and barriers to preservation. The barriers to preservation are explicitly depicted by bold lines in the model. The model has four primary sections: data creation, quality control, content, and data collections.

## Data Creation

In the Digital Research Data Lifecycle, creating data is the first step in the process (see Figure 2).

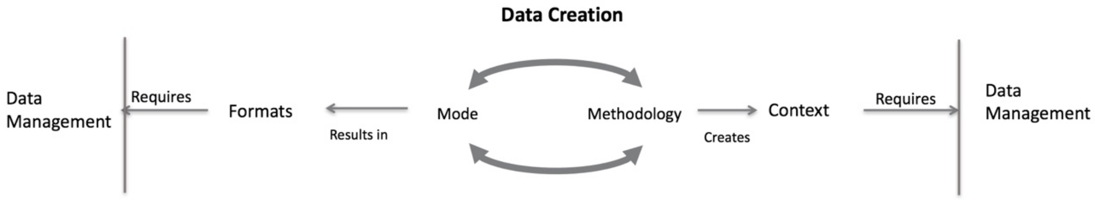

**Figure 2.** The Data Creation Process.

The data creation process has two major components – mode and methodology. The mode is the manner of creation, either generating or gathering. 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 Digital Research Data Lifecycle, 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, modelling and simulation, and text or language analysis.

Of the 796 researchers who responded to the question regarding their data generating practices, 91% created new data and 42% gathered existing data (see Table 1).

**Table 1.** 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% |

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.

Data is highly dependent on the research methodologies; thus, methodologies have a major impact on the Digital Research Data Lifecycle. Participants in this study were asked to identify the research methodologies that they use (see Table 2). Individuals could respond with multiple answers. The 782 responders used a total of 2,059 methodologies (an average of 2.7 responses per individual). The participants used surveys (19%), field studies (29%), case studies (12%), direct observation in experimental situations (46%), analysis of instrument generated data (49%), analysis of existing data sets (38%), modelling and simulation (48%), text or language analysis (8%).

**Table 2.** 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% |

A majority of the respondents (79%) used multiple research methodologies, with 67% using between two and four different methodologies. This opens a new window into the workings of digital research. If each of these methodologies creates multiple file formats, the complexity of the data to curate will grow with each additional methodology.

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 3).

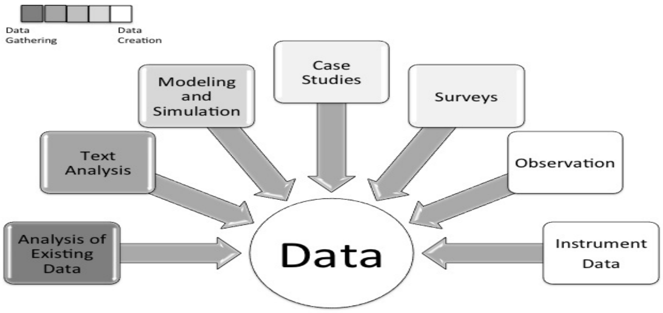

**Figure 3.** Data Creation and Methodology.

Clearly, analysis of existing data relies on gathering data. This is represented in Figure 3 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. Modelling 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.

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 data resulting from statistical and/or text analysis processes 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.

### 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 for SAS (both proprietary). 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 well as community semantic and syntactic standards such as FITS[1] (for astronomical data), NetCDF[2] (for array data), SEG-Y[3] (for seismic data), ROOT[4] (for high energy physics), and FASTA[5] (for protein sequences).

### Data Creation and Context

Contextual information generated during the process of creating research data includes data sources, instrumentation used, instrumentation settings, experimental conditions, software used, software configurations, and samples used. Both data generating and data gathering modes create contextual information. 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.

[1] The FITS Support Office at NASA/GSFC: https://fits.gsfc.nasa.gov/

[2] NetCDF: https://www.unidata.ucar.edu/software/netcdf/docs/

[3] USGS Seismic Data Format: https://pubs.usgs.gov/of/2001/of01-326/HTML/FILEFORM.HTM

[4] ROOT Data Analysis Framework: https://root.cern.ch/about-root

[5] Basic Local Alignment Search Tool (BLAST): https://blast.ncbi.nlm.nih.gov/

### 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, newly created or integrated data, and contextual data.

## 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 and Scannapieco, 2006), is the second step in the Digital Research Data Lifecycle model (see Figure 4).

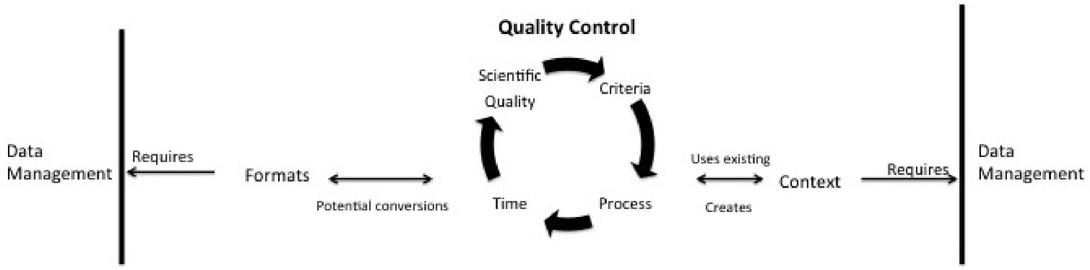

**Figure 4.** Quality Control Process.

Ensuring data quality involves developing and applying processes to combine data from numerous sources, to manipulate data to reconcile different scales of measurement, and to validate the content. 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 (see Table 3). 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 ($F_{4,724} = 8.927, p = .604$). 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.

**Table 3.** 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% |

The researchers were asked about the quality control processes that they implemented (see Table 4). The 711 researchers who answered this question used an average of 2.4 processes per project including data normalization, cleaning, integration, and calibration. Data normalizing (resolving scale issues, reformatting for consistency, etc.) was used by 67% of the responders; data cleaning (fixing errors) was used by 59% of the responders; data integration (merging data from several sources) was used by 63% of responders, and instrument calibration was used by 41% of the responders.

**Table 4.** 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. 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.

Each of the quality control processes was sensitive to the size of lab as well. 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.

Scientific quality and quality of data are tightly coupled. A large majority of researchers (88%) 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.

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. 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.

### 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 opaque, proprietary statistical output file is produced (as an .spv or .spo file). In SPSS version 24, this proprietary output file can be exported in a number of other formats: Excel (.xls), HTML (.htm), Portable Document Format (.pdf), Text (.txt), Microsoft Word/RTF (.doc), and Microsoft PowerPoint (.ppt) (IBM, 2016).

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.

### 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[6] plugin to an XML editor), each of which produces and/or contains contextual data.

[6] Schematron: http://schematron.com/

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

**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. 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 (57%) and/or graduate students (38%), 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.

## Content

Through the research process, data becomes content as it has value, form, and meaning (see Figure 5).

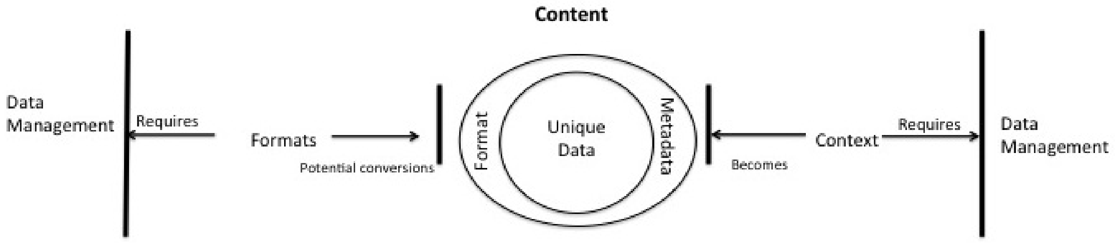

**Figure 5**. Data to Content Process.

For this study, content is defined as the principal, significant intellectual and informational substance of the research process. For data to become content, the scientific process must have added value to the data making it unique in some way that extends human knowledge and that adds to the scientific record. For data to become content, it must have a form that is readable, renderable, and usable; it must be understood within its context; it must have metadata that describes its meaning. As with the other stages in the Digital Research Data Lifecycle, 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.

Uniqueness is an important assessment criterion for preservation and 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 need to be preserved (Gray et al., 2005; Henty et al., 2008; Key Perspectives, 2010; Lord and McDonald, 2003; Lyon, 2007).

The researchers in this study rejected the simplistic binary assessment of uniqueness (unique or not unique) as described in the literature. The researchers were asked to choose from a list of possible options that were developed from the responses from the preliminary study (see Table 5). 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.

**Table 5.** 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% |

Researchers indicate that uniqueness is related to the value of the data which can be attributed to the contributions of the scientific process. Researchers can add value to existing data that makes the data unique; through the research process, data becomes 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. 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. The researchers disagree with the proposition that the processing to create uniformity or integration is simple computation. Much of this work is done by hand, requires intellectual input, and becomes irreplaceable. This data cannot be easily recreated; the processes that create this data cannot be easily rerun.

### Content and format

All throughout the Digital Research Data Lifecycle, 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 (79%). 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 (31%) or are formats that are opaque, are commercially owned, and have strong commercial software dependencies such as Microsoft Office formats for Word or Excel (48%). 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.

### Content and context

In order for data to be transformed into content, its meaning and 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 formats use, as a method to encode the meaning, context, and structure of the data, 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. 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. Only 20% indicated that they would be willing to spend research funds on a metadata professional. The researchers expressed concerns about their ability to use the specialists effectively, sufficient work to keep a full time person occupied, the cost of transferring complex, domain specific knowledge, and the willingness of agencies to fund metadata enhancements.

### 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, 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.

## Data Collections

Making arrangements for the final disposition of data is the last step of the Digital Research Data Lifecycle model. As defined by the Merriam-Webster dictionary (2011), disposition means an orderly arrangement or the transfer of administrative control to another. Within the Digital Research Data Lifecycle, 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 6). 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 areas 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). 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 and Shankar, 2011). Nevertheless, the framework 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. While this typology is useful, its simplicity can mask the complexity of the repository landscape; many repositories could be classified in multiple categories.

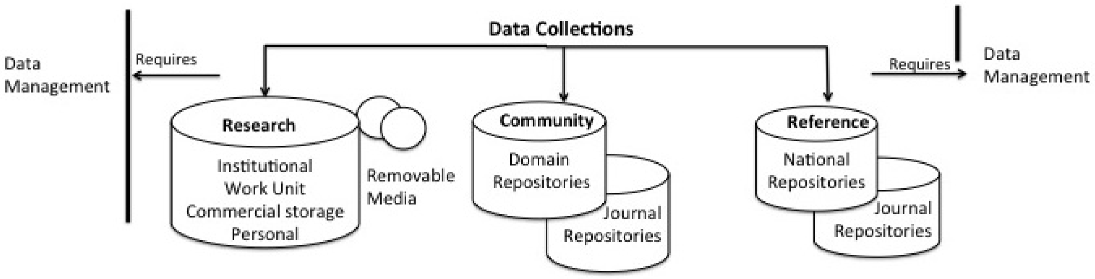

**Figure 6.** Data Collections.

Research collections are by far the most common disposition of data (72%). Research collections are primarily supported by the individual researcher on removable media such as media as CDs, DVDs, or hard drives. 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 (35%). 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 collections. These archives take responsibility for data management, long term preservation, and ongoing access. In this study, 16% of researchers indicate that they use their institutional repository to store their data at the end of a project. A small number of researchers have begun to use commercially available storage services such as Amazon and Google for their research collections. These services may provide a stable technology base that provides online access for a very low cost. However, these cloud services can be based on proprietary systems with opaque management and preservation with dubious terms of service that may cede ownership of stored data/documents to the commercial entity (Instrumental, 2013; National Archives and Records Administration, 2010).

Community (13%) or reference data collections (14%) serve as the final data disposition for a minority of researchers in this study. Researchers have a variety of motivations for contributing their data to these collections including funding agency mandates, (25%) journal mandates (16%), personal initiative (28%), and standard practice in research work group (25%).[7]

### Data collections and format

In the Digital Research Data Lifecycle, format conversions occur throughout the research process. The type of data collection determines the future format conversion scenarios. For research collections, the most likely scenario is stasis; 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.

[7] For a fuller discussion of the final disposition of research data, see Kowalczyk, S.T. (2014).

### 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. For all types of repositories, researchers in this study who used repositories found deposit processes to be easy to use (66%). 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.

### 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 and Shankar, 2011). Despite the opacity of the services, researchers are using these services with the expectation of persistent archiving and preservation.

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.

### 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 with only 6% of respondents 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.

## Discussion

The Digital Research Data Lifecycle is a complex interaction of content, formats, context, quality control, data collections, and the technical infrastructure of the researcher's home institution. It is, like other lifecycles, a visual and conceptual representation of a set of organic processes and stages. Lifecycles are a useful way to depict digital curation as they can describe the process of creating, processing, maintaining, and sharing data. The research behind the various data lifecycles have each focused on specific issues. The lifecycles can represent the repositories' view of the process; or the lifecycle can represent a research domain or the needs of a specific project or lab. Each lifecycle is the response to a specific set of research questions, a specific problem that needs to be addressed, and/or a specific community of users. The Digital Research Data Lifecycle depicts the antecedents and the barriers to preservation within the research process.

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.

### Barriers to Preservation

This research has identified a set of antecedents to preservation. These antecedents – data management, contextual metadata, format, 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. Little of the data created by researchers use syntactic and semantic community or domain data standard formats, making this data more difficult to use and preserve over time. 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.

### 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.

#### 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., 2005; Henty et al., 2008; Key Perspectives, 2010; Lord and 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.

#### File formats and standards

File formats for both research data and contextual metadata are a major component of the Digital Research Data Lifecycle. 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 element 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 can easily become useless. Thus, the lack of syntactic and semantic domain or community based standard file formats is a significant barrier to preservation.

## Conclusion

Curating, preserving, and providing access to scientific data is vital to the health of the scientific enterprise (Iwata, 2008; Rusbridge, 2007). This research used a mixed methodology approach to develop a life cycle model for the preservation of research data that accounts for both the antecedents and barriers to preservation. 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. 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 multiple environments in which data is created, used, saved, and preserved.

## References

Abrams, S.L. (2004). The role of format in digital preservation. VINE, 34(2), 49-55.

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, 191-201.

Association of Research Libraries. (2006). To stand the test of time: Long-term stewardship of digital data sets in science and engineering A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, September 26-27. Arlington, VA: American Research Libraries.

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

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

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. Science, 323(5919), 1297-1298. doi:10.1126/science.1170411

Borgman, C.L. (2007). Scholarship in the digital age. Cambridge, MA: The MIT Press.

Borgman, C.L., Bowker, G.C., Finholt, T.A., & Wallis, J.C. (2009). Towards a virtual organization for data Cyberinfrastructure. Paper presented at the Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, Austin, Texas.

Brown, A. (2006). Automatic format identification using PRONOM and DROID. Digital Preservation Technical Paper 1, Issue 2: National Archives of the United Kingdom.

Carlson, J., Johnston, L., & Huang, Y. (Eds.). (2014). Data information literacy: Librarians, data, and the education of a new generation of researchers. Purdue University Press.

Cheung, K., Hunter, J., Lashtabeg, A., & Drennan, J. (2008). SCOPE-A scientific compound object publishing and editing system. International Journal of Digital Curation, 3(2), 4-18.

Choudhary, A., Kandemir, M.T., No, J., Memik, G., Shen, X., Liao, Nagesh, H., More, S., Taylor, V., Thakur, R., & Stevens, R. (2000). Data management for large-scale scientific computations in high performance distributed systems. Cluster Computing, 3(1), 45-60. doi:10.1023/A:1019063700437

Clements, A., & McCutcheon, V. (2014). Research data meets research information management: Two case studies using (a) Pure CERIF-CRIS and (b) EPrints repository platform with CERIF extensions. Procedia Computer Science, 33, 199-206.

Coles, S., Carr, L., & Frey, J. (2007). The Repository for the Laboratory (R4L): Final report. Southampton, UK: Joint Information Systems Committee (JISC) Digital Repositories Programme and the University of Southampton.

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.

Crow, R. (2002). The case for institutional repositories: A SPARC position paper (Release 1.0). Washington, DC: The Scholarly Publishing and Academic Resources Coalition.

Disposition. (2011) Merriam-Webster Dictionary. Merriam-Webster, Incorporated.

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

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

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. doi:10.1145/1107499.1107503

Hacker, T.J., & Wheeler, B.C. (2007). Making research cyberinfrastructure a strategic choice. EDUCAUSE Quarterly, 30(1), 21-29.

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).

Henty, M., Weaver, B., Bradbury, S., & Porter, S. (2008). Investigating data management practices in Australian universities. Canberra, Australia: Australian Partnership for Sustainable Repositories.

Hey, T. (2010). Data-intensive scientific discovery: The fourth paradigm. Digital Science Center Seminar Series, Pervasive Technology Institute, Indiana University. Bloomington, Indiana.

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 Infrastrucutre a Reality (pp. 809-824). Chichester, UK: John Wiley & Sons, Ltd.

Higgins, S. (2008). The DCC curation lifecycle model. International Journal of Digital Curation, 13(1), 134-140.

Humphrey, C. (2006). e-Science and the life cycle of research. University of Alberta, Canada.

IBM. (2016). IBM SPSS Statistics 24 Brief Guide. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Statistics_Brief_Guide.pdf

Instrumental. (2013). Report on digital preservation and cloud services. St. Paul, MN: Minnesota Historical Society. Retrieved from http://www.mnhs.org/preserve/records/docs_pdfs/Instrumental_MHSReportFinal_Public_v2.pdf

Iwata, S. (2008). Editor's note: Scientific 'agenda' of data science. Data Science Journal, 7, 54-56.

Jones, S., Ball, A., & Ekmekcioglu, Ç. (2008). The data audit framework: A first step in the data management challenge. International Journal of Digital Curation, 2(3).

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

Key Perspectives Ltd. (2010). Data dimensions: Disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study. Edinburgh, UK: Digital Curation Centre.

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

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. (2014). Where does all the data go: Quantifying the final disposition of research data. Proceedings of the Association for Information Science and Technology, 51(1), 1-10. doi:10.1002/meet.2014.14505101044

LaRowe, G., Ambre, S., Burgoon, J., Ke, W., & Börner, K. (2009). The scholarly database and its utility for scientometrics research. Scientometrics, 29(2), 219-234. doi:10.1007/s11192-009-0414-2

Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., & Kenney, A.R. (2000). Risk management of digital information: A file format investigation. Washington, D.C.: Council on Library and Information Resources.

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

Library of Congress. (2011). Preserving our digital heritage: The national digital information infrastructure and preservation program 2010 report. Washington, DC: Library of Congress.

Long, D.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. Paper presented at the 40th IEEE Computer Society International Conference (COMPCON '95), San Francisco, CA, March 05-09.

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.

Lynch, C. (2008). Big data: How do your data grow? Nature, 455(7209), 28-29. doi:10.1038/455028a

Lyon, L. (2007). Dealing with data: Roles, rights, responsibilities and relationships. Consultancy report. Bath, UK: UKOLN and Joint Information Systems Committee (JISC).

Mann, R., Williams, R., Atkinson, M., Brodlie, K., Storkey, A., & Williams, C. (2002). Scientific data mining, integration, and visualization: Report of the workshop held at the e-Science Institute

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. Minneapolis, MN: University of Minnesota.

Mayernik, M.S. (2015). Research data and metadata curation as institutional issues. Journal of the Association for Information Science and Technology.

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

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

Minor, D., Critchlow, M., Hutt, A., Fleming, D., Bergstrom, M.L., & Sutton, D. (2014). Research data curation pilots: Lessons learned. International Journal of Digital Curation, 9(1), 220-230.

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

National Archives and Records Administration [NARA]. (2010). Guidance on managing records in cloud computing environments: NARA Bulletin 2010-05. Retrieved from http://www.archives.gov/records-mgmt/bulletins/2010/2010-05.html

NISO Framework Working Group (NISO). (2007). A framework of guidance for building good digital collections. NISO Baltimore, Maryland. Retrieved from https://www.niso.org/sites/default/files/2017-08/framework3.pdf

National Science Board. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. Washington, DC: National Science Board Committee on Programs and Plans, NSB-05-40.

Pinfield, S., Cox, A.M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. PloS ONE, 9(12), e114734.

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.

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.

Rajasekar, A.K., & Moore, R.W. (2001). Data and metadata collections for scientific applications. Paper presented at the Proceedings of the 9th International Conference on High-Performance Computing and Networking (HPCN Europe 2001), Amsterdam, The Netherlands.

Ray, J.M. (2014). Research data management: Practical strategies for information professionals. Purdue University Press.

Rice, R. (2007). DISC-UK DataShare projects: Building exemplars for institutional data repositories in the UK. iASSIST(Fall/Winter), 21-27.

Ross, S. (2007). Digital preservation, archival science and methodological foundations for digital libraries. Paper presented at the Keynote Address, 11th European Conference on Digital Libraries (ECDL 2007), Budapest, Hungary.

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).

Rusbridge, C. (2007). Create, curate, re-use: The expanding life course of digital research data. Paper presented at the Proceedings of EDUCAUSE Australasia 2007: Advancing Knowledge, Pushing Boundaries, Melbourne, Australia.

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

Steinhart, G. (2007). DataSTaR: An institutional approach to research data curation. iASSIST Quarterly, 31(3-4), 34-39.

Strauss, A.L., & Corbin, J.M. (1990). Basics of qualitative research: Grounded theory procedures and techniques. 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, JISC and the National Environment Research Council UK.

Tibbo, H.R. (2014). Digital and data curation. Retrieved from http://cihe.skku.edu/download/Digital_and_Data_Curation-black.pdf

Treloar, A., Groenewegen, D., & Harboe-Ree, C. (2007). The data curation continuum: Managing data objects in institutional repositories. D-Lib Magazine, 13(9/10).

Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., et al. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS ONE 10(8): e0134826. doi:10.1371/journal.pone.0134826

UK Data Service. (2018). Research data lifecycle. Retrieved from https://www.ukdataservice.ac.uk/manage-data/lifecycle

van Westrienen, G., & Lynch, C.A. (2005). Academic institutional repositories: Deployment status in 13 nations as of mid-2005. D-Lib Magazine, 11(9).

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.

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), Article No. 3.

Vines, T.H., Albert, A.Y., Andrew, R.L., Débarre, F., Bock, D.G., Franklin, M.T., ... & Rennison, D.J. (2014). The availability of research data declines rapidly with article age. Current Biology, 24(1), 94-97.

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

Wallis, J.C., Borgman, C.L., Mayernik, M.S., & 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, 3(1), 114-126.

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

Wilkes, W., Brunsmann, J., Heutelbeck, D., Hundsdörfer, A., Hemmje, M., & Heidbrink, H.U. (2011). Towards support for long-term digital preservation in product life cycle management. International Journal of Digital Curation, 6(1), 282-296.

Witt, M., Carlson, J., Brandt, D.S., & Cragin, M.H. (2009). Constructing data curation profiles. International Journal of Digital Curation, 4(3), 93-103.

Zilinski, L., Scherer, D., Bullock, D., Horton, D., & Matthews, C. (2014). Evolution of data creation, management, publication, and curation in the research process. Transportation Research Board 93rd Annual Meeting, Washington, DC.